Statistica Sinica **7**(1997), 221-264

AN ASYMPTOTIC THEORY FOR LINEAR MODEL SELECTION

Jun Shao

University of Wisconsin

Abstract: In the problem of selecting a linear model to approximate the true unknown regression model, some necessary and/or sufficient conditions are established for the asymptotic validity of various model selection procedures such as Akaike's AIC, Mallows' C_p, Shibata's FPE_λ, Schwarz' BIC, generalized AIC, cross-validation, and generalized cross-validation. It is found that these selection procedures can be classified into three classes according to their asymptotic behavior. Under some fairly weak conditions, the selection procedures in one class are asymptotically valid if there exist fixed-dimension correct models; the selection procedures in another class are asymptotically valid if no fixed-dimension correct model exists. The procedures in the third class are compromises of the procedures in the first two classes. Some empirical results are also presented.

Key words and phrases: AIC, asymptotic loss efficiency, BIC, consistency, C_p, cross-validation, GIC, squared error loss.

1. Introduction

Let $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$ be a vector of n independent responses and $\boldsymbol{X}_n = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$ be an $n \times p_n$ matrix whose ith row $\boldsymbol{x}_i$ is the value of a p_n-vector of explanatory variables associated with y_i. For inference purposes, a class of models, indexed by $\alpha \in \mathcal{A}_n$, is to characterize the relation between the mean response $\boldsymbol{\mu}_n = E(\boldsymbol{y}_n|\boldsymbol{X}_n)$ and the explanatory variables. If $\mathcal{A}_n$ contains more than one model, then we need to select a model from $\mathcal{A}_n$ using the given $\boldsymbol{X}_n$ and the data vector $\boldsymbol{y}_n$. The following are some typical examples.

Example 1. Linear regression. Suppose that $p_n = p$ for all n and $\boldsymbol{\mu}_n = \boldsymbol{X}_n\boldsymbol{\beta}$ with an unknown p-vector $\boldsymbol{\beta}$. Write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ and $\boldsymbol{X}_n = (\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})$. It is suspected that the sub-vector $\boldsymbol{\beta}_2 = \mathbf{0}$, i.e., $\boldsymbol{X}_{n2}$ is actually not related to $\boldsymbol{\mu}_n$. Then we may propose the following two models:

$$\begin{array}{ll}\text{Model 1:} & \boldsymbol{\mu}_n = \boldsymbol{X}_{n1}\boldsymbol{\beta}_1 \\ \text{Model 2:} & \boldsymbol{\mu}_n = \boldsymbol{X}_n\boldsymbol{\beta}\end{array}.$$

In this case, $\mathcal{A}_n = \{1, 2\}$. It is well known that the least squares fitting under model 1 is more efficient than that under model 2 if $\boldsymbol{\beta}_2 = \mathbf{0}$. More generally, we can consider models

$$\boldsymbol{\mu}_n = \boldsymbol{X}_n(\alpha)\boldsymbol{\beta}(\alpha), \tag{1.1}$$

where α is a subset of $\{1,\ldots,p\}$ and $\boldsymbol{\beta}(\alpha)$ (or $\boldsymbol{X}_n(\alpha)$) contains the components of $\boldsymbol{\beta}$ (or columns of $\boldsymbol{X}_n$) that are indexed by the integers in α. In this case $\mathcal{A}_n$ consists of some distinct subsets of $\{1,\ldots,p\}$. If $\mathcal{A}_n$ contains all nonempty subsets of $\{1,\ldots,p\}$, then the number of models in $\mathcal{A}_n$ is 2^p-1.

Example 2. One-mean versus k-mean. Suppose that n observations are from k groups. Each group has r observations that are identically distributed. Thus, $n=kr$, where $k=k_n$ and $r=r_n$ are integers. Here we need to select one model from the following two models: (1) the one-mean model, i.e., the k groups have a common mean; (2) the k-mean model, i.e., the k groups have different means. To use the same formula as that in (1.1), we define $p_n=k$,

$$\boldsymbol{X}_n=\begin{pmatrix} \mathbf{1}_r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_r & \mathbf{1}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_r & \mathbf{0} & \mathbf{1}_r & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{1}_r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_r \end{pmatrix}$$

and $\boldsymbol{\beta}=(\mu_1,\mu_2-\mu_1,\ldots,\mu_k-\mu_1)'$, where $\mathbf{1}_r$ denotes the r-vector of ones. Then $\mathcal{A}_n=\{\alpha_1,\alpha_k\}$, where $\alpha_1=\{1\}$ and $\alpha_k=\{1,\ldots,k\}$.

Example 3. Linear approximations to a response surface. Suppose that we wish to select the best approximation to the true mean response surface from a class of linear models. Note that the approximation is exact if the response surface is actually linear and is in $\mathcal{A}_n$. The proposed models are $\boldsymbol{\mu}_n=\boldsymbol{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha),\ \alpha\in\mathcal{A}_n$, where $\boldsymbol{X}_n(\alpha)$ is a sub-matrix of $\boldsymbol{X}_n$ and $\boldsymbol{\beta}_n(\alpha)$ is a sub-vector of a p_n-vector $\boldsymbol{\beta}_n$ whose components have to be estimated. As a more specific example, we consider the situation where we try to approximate a one-dimensional curve by a polynomial, i.e., $\boldsymbol{\mu}_n=\boldsymbol{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha)$ with the ith row of $\boldsymbol{X}_n(\alpha)$ being $(1,t_i,t_i^2,\ldots,t_i^{h-1})'$, $i=1,\ldots,n$. In this case $\mathcal{A}_n=\{\alpha_h,h=1,\ldots,p_n\}$ and $\alpha_h=\{1,\ldots,h\}$ corresponds to a polynomial of order h used to approximate the true model. The largest possible order of the polynomial may increase as n increases, since the more data we have, the more terms we can afford to use in the polynomial approximation.

We assume in this paper that the models in $\mathcal{A}_n$ are linear models and the least squares fitting is used under each proposed model. Each model in $\mathcal{A}_n$ is denoted by α, a subset of $\{1,\ldots,p_n\}$. After observing the vector $\boldsymbol{y}_n$, our concern is to select a model α from $\mathcal{A}_n$ so that the squared error loss

$$L_n(\alpha)=\frac{\|\boldsymbol{\mu}_n-\hat{\boldsymbol{\mu}}_n(\alpha)\|^2}{n} \tag{1.2}$$

be as small as possible, where $\| \ \|$ is the Euclidean norm and $\hat{\boldsymbol{\mu}}_n(\alpha)$ is the least squares estimator (LSE) of $\boldsymbol{\mu}_n$ under model α. Note that minimizing $L_n(\alpha)$ is equivalent to minimizing the average prediction error $E[n^{-1}\|\boldsymbol{z}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\|^2|\boldsymbol{y}_n]$, where $\boldsymbol{z}_n = (z_1, \ldots, z_n)'$ and z_i is a future observation associated with $\boldsymbol{x}_i$ and is independent of y_i.

A considerable number of selection procedures were proposed in the literature, e.g., the AIC method (Akaike (1970)); the C_p method (Mallows (1973)); the BIC method (Schwarz (1978); Hannan and Quinn (1979)); the FPE_λ method (Shibata (1984)); the generalized AIC such as the GIC method (Nishii (1984), Rao and Wu (1989)) and its analogues (Pötscher (1989)); the delete-1 cross-validation (CV) method (Allen (1974), Stone (1974)); the generalized CV (GCV) method (Craven and Wahba (1979)); the delete-d CV method (Geisser (1975), Burman (1989), Shao (1993), Zhang (1993)); and the PMDL and PLS methods (Rissanen (1986), Wei (1992)). Some asymptotic results in assessing these selection procedures have been established in some particular situations. Nishii (1984) and Rao and Wu (1989) showed that in Example 1, the BIC and GIC are consistent (definitions of consistency will be given in Section 2), whereas the AIC and C_p are inconsistent. On the other hand, Stone (1979) showed that in some situations (Example 2), the BIC is inconsistent but the AIC and C_p are consistent. In Example 3, Shibata (1981) and Li (1987) showed that the AIC, the C_p, and the delete-1 CV are asymptotically correct in some sense. However, Shao (1993) showed that in Example 1, the delete-1 CV is inconsistent and the delete-d CV is consistent, provided that $d/n \to 1$. These results do not provide a clear picture of the performance of the various selection procedures. Some of these conclusions are obviously contrary to each other. But this is because these results are obtained in quite different circumstances. A crucial factor that almost determines the asymptotic performances of various model selection procedures is whether or not $\mathcal{A}_n$ contains some correct models in which the dimensions of regression parameter vectors do not increase with n. This will be explored in detail in the current paper.

The purpose of this paper is to provide an asymptotic theory which shows when the various selection procedures are asymptotically correct (or incorrect) under an asymptotic framework covering all situations described in Examples 1-3. After introducing some notations and definitions in Section 2, we study the asymptotic behavior of the GIC method in Section 3 and other selection procedures cited above in Section 4. Some numerical examples are given in Section 5. Section 6 contains some technical details.

2. Notation and Definitions

Throughout the paper we assume that $(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}$ exists and that the minimum and maximum eigenvalues of $\boldsymbol{X}_n'\boldsymbol{X}_n$ are of order n. The matrices $\boldsymbol{X}_n$, $n = 1, 2, \ldots$, are considered to be non-random. The results in this paper are also valid in the almost sure sense when the $\boldsymbol{X}_n$ are random, provided that the required conditions involving $\boldsymbol{X}_n$ hold for almost all sequences $\boldsymbol{X}_n$, $n = 1, 2, \ldots$

Let $\mathcal{A}_n$ be a class of proposed models (subsets of $\{1, \ldots, p_n\}$) for selection. The number of models in $\mathcal{A}_n$ is finite, but may depend on n. For $\alpha \in \mathcal{A}_n$, the proposed model is $\boldsymbol{\mu}_n = \boldsymbol{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha)$, where $\boldsymbol{X}_n(\alpha)$ is an $n \times p_n(\alpha)$ submatrix of the $n \times p_n$ matrix $\boldsymbol{X}_n$ and $\boldsymbol{\beta}_n(\alpha)$ is a $p_n(\alpha) \times 1$ sub-vector of an unknown $p_n \times 1$ vector $\boldsymbol{\beta}_n$. Without loss of generality, we assume that the largest model $\bar{\alpha}_n = \{1, \ldots, p_n\}$ is always in $\mathcal{A}_n$. The dimension of $\boldsymbol{\beta}_n(\alpha)$, $p_n(\alpha)$, will be called the dimension of the model α. Under model α, the LSE of $\boldsymbol{\mu}_n$ is $\hat{\boldsymbol{\mu}}_n(\alpha) = \boldsymbol{H}_n(\alpha)\boldsymbol{y}_n$, where $\boldsymbol{H}_n(\alpha) = \boldsymbol{X}_n(\alpha)[\boldsymbol{X}_n(\alpha)'\boldsymbol{X}_n(\alpha)]^{-1}\boldsymbol{X}_n(\alpha)'$.

A proposed model $\alpha \in \mathcal{A}_n$ is said to be *correct* if $\boldsymbol{\mu}_n = \boldsymbol{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha)$ is actually true. Note that $\mathcal{A}_n$ may not contain a correct model (Example 3); a correct model is not necessarily the best model, since there may be several correct models in $\mathcal{A}_n$ (Examples 1 and 2) and there may be an incorrect model having a smaller loss than the best correct model (Example 2). Let

$$\mathcal{A}_n^c = \{\alpha \in \mathcal{A}_n : \boldsymbol{\mu}_n = \boldsymbol{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha)\}$$

denote all the proposed models that are actually correct models. It is possible that $\mathcal{A}_n^c$ is empty or $\mathcal{A}_n^c = \mathcal{A}_n$.

Let $\boldsymbol{e}_n = \boldsymbol{y}_n - \boldsymbol{\mu}_n$. It is assumed that the components of $\boldsymbol{e}_n$ are independent and identically distributed with $V(\boldsymbol{e}_n|\boldsymbol{X}_n) = \sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the identity matrix of order n. The loss defined in (1.2) is equal to $L_n(\alpha) = \Delta_n(\alpha) + (\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n)/n$, where $\Delta_n(\alpha) = (\|\boldsymbol{\mu}_n - \boldsymbol{H}_n(\alpha)\boldsymbol{\mu}_n\|^2)/n$. Note that $\Delta_n(\alpha) = 0$ when $\alpha \in \mathcal{A}_n^c$. The risk (the expected average squared error) is

$$R_n(\alpha) = E[L_n(\alpha)] = \Delta_n(\alpha) + \frac{\sigma^2 p_n(\alpha)}{n}.$$

Let $\hat{\alpha}_n$ denote the model selected using a given selection procedure and let α_n^L be a model minimizing $L_n(\alpha)$ over $\alpha \in \mathcal{A}_n$. The selection procedure is said to be *consistent* if

$$P\left\{\hat{\alpha}_n = \alpha_n^L\right\} \to 1 \tag{2.1}$$

(all limiting processes are understood to be as $n \to \infty$). Note that (2.1) implies

$$P\left\{L_n(\hat{\alpha}_n) = L_n(\alpha_n^L)\right\} \to 1. \tag{2.2}$$

Thus, $\hat{\boldsymbol{\mu}}_n(\hat{\alpha}_n)$ is in this sense asymptotically as efficient as the best estimator among $\hat{\boldsymbol{\mu}}_n(\alpha)$, $\alpha \in \mathcal{A}_n$. (2.1) and (2.2) are equivalent if $L_n(\alpha)$ has a unique minimum for all large n.

The consistency defined in (2.1) is in terms of model selection, i.e., we treat $\hat{\alpha}_n$ as an "estimator" of α_n^L (it is a well defined estimator if α_n^L is non-random, e.g., in Example 1). This consistency is not related to the consistency of $\hat{\boldsymbol{\mu}}_n(\hat{\alpha}_n)$ as an estimator of $\boldsymbol{\mu}_n$, i.e., $L_n(\hat{\alpha}_n) \to_p 0$. In fact, it may not be worthwhile to discuss the consistency of $\hat{\boldsymbol{\mu}}_n(\hat{\alpha}_n)$, since sometimes there is no consistent estimator of $\boldsymbol{\mu}_n$ (e.g., $\min_{\alpha \in \mathcal{A}_n} L_n(\alpha) \not\to_p 0$) and sometimes there are too many consistent estimators of $\boldsymbol{\mu}_n$ (e.g., $\max_{\alpha \in \mathcal{A}_n} L_n(\alpha) \to_p 0$, in which case $\hat{\boldsymbol{\mu}}_n(\alpha)$ is consistent for any α).

In some cases a selection procedure does not have property (2.1), but $\hat{\alpha}_n$ is still "close" to α_n^L in the following sense that is weaker than (2.1):

$$L_n(\hat{\alpha}_n)/L_n(\alpha_n^L) \to_p 1, \tag{2.3}$$

where $\to_p$ denotes convergence in probability. A selection procedure satisfying (2.3) is said to be *asymptotically loss efficient*, i.e., $\hat{\alpha}_n$ is asymptotically as efficient as α_n^L in terms of the loss $L_n(\alpha)$. Since the purpose of model selection is to minimize the loss $L_n(\alpha)$, (2.3) is an essential asymptotic requirement for a selection procedure.

Clearly, consistency in the sense of (2.1) implies asymptotic loss efficiency in the sense of (2.3). In some cases (e.g., Examples 1 and 2), consistency is the same as asymptotic loss efficiency. The proof of the following result is given in Section 6.

Proposition 1. *Suppose that*

$$p_n/n \to 0, \tag{2.4}$$

$$\liminf_{n\to\infty} \min_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \Delta_n(\alpha) > 0 \tag{2.5}$$

and $\mathcal{A}_n^c$ is nonempty for sufficiently large n. Then (2.1) *is equivalent to* (2.3) *if either $p_n(\alpha_n^L) \not\to_p \infty$ or $\mathcal{A}_n^c$ contains exactly one model for sufficiently large n.*

The following regularity condition will often be used in establishing asymptotic results:

$$\sum_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{1}{[nR_n(\alpha)]^l} \to 0, \tag{2.6}$$

where l is some fixed positive integer such that $E(y_1 - \mu_1)^{4l} < \infty$. Note that condition (2.6) is exactly the same as condition (A.3) in Li (1987) when $\mathcal{A}_n^c$ is empty; but Li's condition (A.3) may not hold when $\mathcal{A}_n^c$ is not empty. If the number of models in $\mathcal{A}_n$ is bounded (Examples 1 and 2), then (2.6) with $l = 1$ is the same as

$$\min_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} nR_n(\alpha) \to \infty, \tag{2.7}$$

the condition (A.3′) in Li (1987). When $\mathcal{A}_n = \{\alpha_h, h = 1, \ldots, p_n\}$ with $\alpha_h = \{1, \ldots, h\}$ (e.g., polynomial approximation in Example 3), Li (1987) showed that condition (2.6) with $l = 2$ is the same as (2.7). Under an additional assumption that $\boldsymbol{e}_n$ is normal, we may replace (2.6) by $\sum_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \delta^{nR_n(\alpha)} \to 0$ for any $0 < \delta < 1$, which is Assumption 2 in Shibata (1981).

3. The GIC$_{\lambda_n}$ Method

Many model selection procedures are identical or equivalent to the procedure which minimizes

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{\lambda_n \hat{\sigma}_n^2 p_n(\alpha)}{n} \tag{3.1}$$

over $\alpha \in \mathcal{A}_n$, where $S_n(\alpha) = \|\boldsymbol{y}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\|^2$, $\hat{\sigma}_n^2$ is an estimator of σ^2, and $\{\lambda_n\}$ is a sequence of non-random numbers ≥ 2 and $\lambda_n/n \to 0$. This procedure will be called the GIC$_{\lambda_n}$ method. If $\hat{\sigma}_n^2 = S_n(\bar{\alpha}_n)/(n - p_n)$, $\bar{\alpha}_n = \{1, \ldots, p_n\}$, then the GIC$_{\lambda_n}$ with $\lambda_n \to \infty$ is the GIC method in Rao and Wu (1989); the GIC$_{\lambda_n}$ with $\lambda_n \equiv 2$ is the C$_p$ method in Mallows (1973); and the GIC$_{\lambda_n}$ with $\lambda_n \equiv \lambda > 2$ is the FPE$_\lambda$ method in Shibata (1984).

Since the GIC$_{\lambda_n}$ is a good representative of the model selection procedures cited in Section 1, we first study its asymptotic behavior. Let the model selected by minimizing $\Gamma_{n,\lambda_n}(\alpha)$ be $\hat{\alpha}_{n,\lambda_n}$.

Consider first the case of $\lambda_n \equiv 2$. Assume that $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. It is shown in Section 6 that

$$\Gamma_{n,2}(\alpha) = \begin{cases} \frac{\|\boldsymbol{e}_n\|^2}{n} + \frac{2\hat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha) \boldsymbol{e}_n}{n} & \alpha \in \mathcal{A}_n^c \\ \\ \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + o_p(L_n(\alpha)) & \alpha \in \mathcal{A}_n - \mathcal{A}_n^c, \end{cases} \tag{3.2}$$

where the equality for the case of $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$ holds under condition (2.6) and the o_p is uniformly in $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$. It follows directly from (3.2) that $\hat{\alpha}_{n,2}$ is asymptotically loss efficient in the sense of (2.3) if there is no correct model in $\mathcal{A}_n$, i.e., $\mathcal{A}_n^c$ is empty. If $\mathcal{A}_n^c$ is not empty but contains exactly one model for each n, say $\mathcal{A}_n^c = \{\alpha_n^c\}$, then $\hat{\alpha}_{n,2}$ is also asymptotically loss efficient. This can be shown by using (3.2) as follows. If $p_n(\alpha_n^c) \to \infty$, then

$$\frac{2\hat{\sigma}_n^2 p_n(\alpha_n^c)}{n} - \frac{\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha_n^c) \boldsymbol{e}_n}{n} = \frac{\hat{\sigma}_n^2 p_n(\alpha_n^c)}{n} + o_p\left(\frac{\hat{\sigma}_n^2 p_n(\alpha_n^c)}{n}\right) = L_n(\alpha_n^c) + o_p(\alpha_n^c),$$

which, together with (3.2), implies that

$$\Gamma_{n,2}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + o_p\left(L_n(\alpha)\right)$$

uniformly in $\alpha \in \mathcal{A}_n$ and, therefore, $\hat{\alpha}_{n,2}$ is asymptotically loss efficient. If $p_n(\alpha_n^c)$ is fixed, then (2.5) holds (Nishii (1984)), which implies that $\alpha_n^L = \alpha_n^c$ and $\min_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \Gamma_{n,2}(\alpha) \not\to_p 0$ and, therefore, $P\{\hat{\alpha}_{n,2} = \alpha_n^L\} \to 1$, i.e., $\hat{\alpha}_{n,2}$ is consistent in the sense of (2.1).

As the following example indicates, however, $\hat{\alpha}_{n,2}$ may not be an asymptotically loss efficient procedure when $\mathcal{A}_n^c$ contains more than one model.

Example 4. Suppose that $\mathcal{A}_n = \mathcal{A}_n^c = \{\alpha_{n1}, \alpha_{n2}\}$, i.e., $\mathcal{A}_n$ contains two models and both models are correct. Assume that $\alpha_{n1} \subset \alpha_{n2}$. Let p_{n1} and p_{n2} be the dimensions of the models α_{n1} and α_{n2}, respectively. Then $p_{n1} < p_{n2}$ and $\boldsymbol{Q}_n = \boldsymbol{H}_n(\alpha_{n2}) - \boldsymbol{H}_n(\alpha_{n1})$ is a projection matrix of rank $p_{n2} - p_{n1}$. Since $S_n(\alpha_{ni}) = \boldsymbol{e}_n'\boldsymbol{e}_n - \boldsymbol{e}_n'\boldsymbol{H}_n(\alpha_{ni})\boldsymbol{e}_n$, $\hat{\alpha}_{n,2} = \alpha_{n1}$ if and only if $2\hat{\sigma}_n^2(p_{n2} - p_{n1}) > \boldsymbol{e}_n'\boldsymbol{Q}_n\boldsymbol{e}_n$.

Case 1. $p_{n1} \to \infty$. If $p_{n2} - p_{n1} \to \infty$, then $\boldsymbol{e}_n'\boldsymbol{Q}_n\boldsymbol{e}_n/(p_{n2} - p_{n1}) \to_p \sigma^2$ and $P\{\hat{\alpha}_{n,2} = \alpha_{n1}\} \to 1$, i.e., the $\hat{\alpha}_{n,2}$ is consistent. If $p_{n2} - p_{n1} \le q$ for a fixed positive integer q, then $p_{n2}/p_{n1} \to 1$, in which case $L_n(\alpha_{n2})/L_n(\alpha_{n1}) \to_p 1$, i.e., any selection procedure is asymptotically loss efficient.

Case 2. $p_{n1} \not\to \infty$. If $p_{n2} - p_{n1} \to \infty$, then we still have $\boldsymbol{e}_n'\boldsymbol{Q}_n\boldsymbol{e}_n/(p_{n2} - p_{n1}) \to_p \sigma^2$, which implies that $\hat{\alpha}_{n,2}$ is consistent. Assume that $p_{n2} - p_{n1} \not\to \infty$ and that for any fixed integer q and constant $c > 2$,

$$\liminf_{n\to\infty} \inf_{\boldsymbol{Q}_n \in \mathcal{Q}_{n,q}} P\left(\boldsymbol{e}_n'\boldsymbol{Q}_n\boldsymbol{e}_n > c\sigma^2 q\right) > 0, \tag{3.3}$$

where $\mathcal{Q}_{n,q} = \{\text{all } n \times n \text{ projection matrices of rank } q\}$. Note that condition (3.3) holds if $\boldsymbol{e}_n \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. From (3.3) and the fact that $p_{n1} \not\to \infty$ and $p_{n2} - p_{n1} \not\to \infty$, the ratio

$$\frac{L_n(\hat{\alpha}_{n,2})}{L_n(\alpha_{n1})} = I(\hat{\alpha}_{n,2} = \alpha_{n1}) + \frac{L_n(\alpha_{n2})}{L_n(\alpha_{n1})} I(\hat{\alpha}_{n,2} = \alpha_{n2}) = 1 + W_n I(\hat{\alpha}_{n,2} = \alpha_{n2})$$

does not tend to 1, where $W_n = \boldsymbol{e}_n'\boldsymbol{Q}_n\boldsymbol{e}_n/\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha_{n1})\boldsymbol{e}_n$ and $I(C)$ is the indicator function of the set C. For example, when $\boldsymbol{e}_n \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, then $p_{n1}W_n/(p_{n2} - p_{n1})$ is an F-random variable with degrees of freedom $p_{n2} - p_{n1}$ and p_{n1}. Hence $\hat{\alpha}_{n,2}$ is not asymptotically loss efficient.

In Example 4, $\hat{\alpha}_{n,2}$ is asymptotically loss efficient if and only if $\mathcal{A}_n^c$ does not contain two models with fixed dimensions. This is actually true in general. Let α_n^c be the model in $\mathcal{A}_n^c$ with the smallest dimension.

Theorem 1. *Suppose that* (2.6) *holds and that $\hat{\sigma}_n^2$ is consistent for σ^2.*
(i) *If $\mathcal{A}_n^c$ contains at most one model for all n, then $\hat{\alpha}_{n,2}$ is asymptotically loss efficient in the sense of* (2.3)*. Furthermore, if $\mathcal{A}_n^c$ contains a unique model with fixed dimension for all n, then $\hat{\alpha}_{n,2}$ is consistent in the sense of* (2.1)*.*
(ii) *Assume that $\mathcal{A}_n^c$ contains more than one models for sufficiently large n. If*

$$\sum_{\alpha\in\mathcal{A}_n^c}\frac{1}{[p_n(\alpha)]^m}\to 0 \tag{3.4}$$

for some positive integer m such that $E(y_1-\mu_1)^{4m}<\infty$, then $\hat{\alpha}_{n,2}$ is asymptotically loss efficient. If (3.4) *does not hold but*

$$\sum_{\alpha\in\mathcal{A}_n^c,\alpha\neq\alpha_n^c}\frac{1}{[p_n(\alpha)-p_n(\alpha_n^c)]^m}\to 0 \tag{3.5}$$

for some positive integer m such that $E(y_1-\mu_1)^{4m}<\infty$, then $\hat{\alpha}_{n,2}$ is asymptotically loss efficient.
(iii) *Assume that $\mathcal{A}_n^c$ contains more than one models for sufficiently large n and that* (3.3) *holds. Then a necessary condition for $\hat{\alpha}_{n,2}$ being asymptotically loss efficient is that*

$$p_n(\alpha_n^c)\to\infty \quad \textit{or} \quad \min_{\alpha\in\mathcal{A}_n^c,\alpha\neq\alpha_n^c} p_n(\alpha)-p_n(\alpha_n^c)\to\infty. \tag{3.6}$$

(iv) *If the number of models in $\mathcal{A}_n^c$ is bounded, or if $m=2$ and $\mathcal{A}_n=\{\alpha_i, i=1,\ldots,p_n\}$ with $\alpha_i=\{1,\ldots,i\}$, then condition* (3.6) *is also sufficient for the asymptotic loss efficiency of $\hat{\alpha}_{n,2}$.*

Remark 1. Condition (3.6) means that $\mathcal{A}_n$ does not contain two correct models with fixed dimensions.

Remark 2. In Theorem 1, the estimator $\hat{\sigma}_n^2$ is required to be consistent for σ^2. A popular choice of $\hat{\sigma}_n^2$ is $S(\bar{\alpha}_n)/(n-p_n)$, the sum of squared residuals (under the largest model in $\mathcal{A}_n$) over its degree of freedom. This estimator is consistent if $\mathcal{A}_n^c$ is not empty, but is not necessarily consistent when $\mathcal{A}_n^c$ is empty, i.e., there is no correct model in $\mathcal{A}_n$. We shall further discuss this issue in Section 4. If there are a few replicates at each $\boldsymbol{x}_i$, then we can compute the within-group sample variance for each i and the average of the within-group sample variances is always a consistent estimator of σ^2.

Theorem 1 indicates that asymptotically, the GIC$_{\lambda_n}$ method with $\lambda_n\equiv 2$ can be used to find (1) the best model among incorrect models; (2) the better model between a correct model and an incorrect model; but it is too crude to be useful in distinguishing correct models with fixed dimensions, i.e., it tends to overfit (select a correct model with an unnecessarily large dimension).

From (3.1), $\Gamma_{n,\lambda_n}(\alpha)$ is a sum of two components: $S_n(\alpha)/n$, which measures the goodness of fit of model α, and $\lambda_n\hat{\sigma}_n^2 p_n(\alpha)/n$, which is a penalty on the use of models with large dimensions. In view of the fact that the use of $\lambda_n \equiv 2$ tends to overfit, it is natural to consider a large λ_n in (3.1), i.e., to put a heavy penalty on the use of models with large dimensions.

The reason why $\hat{\alpha}_{n,2}$ may not be asymptotically loss efficient is that the minimizer of

$$\Gamma_{n,2}(\alpha) - \frac{\|\boldsymbol{e}_n\|^2}{n} = \frac{2\hat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{n},$$

which is considered as a function of $\alpha \in \mathcal{A}_n^c$, may not be the same as the minimizer of $L_n(\alpha) = \sigma^2 p_n(\alpha)/n$. What will occur if we use a λ_n that $\to \infty$? Similar to the expansion (3.2), we have

$$\Gamma_{n,\lambda_n}(\alpha) = \begin{cases} \frac{\|\boldsymbol{e}_n\|^2}{n} + \frac{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{n} & \alpha \in \mathcal{A}_n^c \\ \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + \frac{(\lambda_n\hat{\sigma}_n^2 - 2\sigma^2)p_n(\alpha)}{n} + o_p(L_n(\alpha)) & \alpha \in \mathcal{A}_n - \mathcal{A}_n^c, \end{cases} \tag{3.7}$$

where the equality for the case of $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$ holds under condition (2.6). If

$$\max_{\alpha \in \mathcal{A}_n^c} \frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)} \to_p 0, \tag{3.8}$$

then, for $\alpha \in \mathcal{A}_n^c$, $\Gamma_{n,\lambda_n}(\alpha) - \|\boldsymbol{e}_n\|^2/n$ is dominated by the term $\lambda_n\hat{\sigma}_n^2 p_n(\alpha)/n$ which has the same minimizer as $L_n(\alpha) = \boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n/n$. Hence,

$$P\{\hat{\alpha}_{n,\lambda_n} \in \mathcal{A}_n^c \ \text{ but } \ \hat{\alpha}_{n,\lambda_n} \neq \alpha_n^c\} \to 0, \tag{3.9}$$

where $\hat{\alpha}_{n,\lambda_n}$ is the model selected using the GIC$_{\lambda_n}$ and α_n^c is the model in $\mathcal{A}_n^c$ with the smallest dimension. This means that the GIC$_{\lambda_n}$ method picks the best model in $\mathcal{A}_n^c$ as long as (3.8) holds, which is implied by a weak condition

$$\limsup_{n\to\infty} \sum_{\alpha \in \mathcal{A}_n^c} \frac{1}{[p_n(\alpha)]^m} < \infty \tag{3.10}$$

for some positive integer m such that $E(y_1 - \mu_1)^{4m} < \infty$. Note that (3.10) holds if the number of models in $\mathcal{A}_n$ is bounded (Examples 1 and 2) or if $m = 2$ and $\mathcal{A}_n = \{\alpha_i, i = 1, \ldots, p_n\}$ with $\alpha_i = \{1, \ldots, i\}$ (polynomial approximation in Example 3).

For the asymptotic correctness of the GIC_{λ_n} method, the remaining question is whether it can assess the models in $\mathcal{A}_n - \mathcal{A}_n^c$. Unfortunately, the GIC_{λ_n} tends to select a model with a small dimension and, therefore, may fail to be asymptotically loss efficient if models with small dimensions have large values of $L_n(\alpha)$. More precisely, if there are α_{n1} and α_{n2} in $\mathcal{A}_n - \mathcal{A}_n^c$ such that

$$\lim_{n\to\infty} \frac{L_n(\alpha_{n1})}{L_n(\alpha_{n2})} > 1 \quad \text{but} \quad \lim_{n\to\infty} \frac{L_n(\alpha_{n1}) + (\lambda_n \hat{\sigma}_n^2 - 2\sigma^2)p_n(\alpha_{n1})/n}{L_n(\alpha_{n2}) + (\lambda_n \hat{\sigma}_n^2 - 2\sigma^2)p_n(\alpha_{n2})/n} < 1 \quad (3.11)$$

(which implies $\lim_{n\to\infty} p_n(\alpha_{n1})/p_n(\alpha_{n2}) < 1$), then the GIC_{λ_n} is not asymptotically loss efficient.

A necessary condition for $\hat{\alpha}_{n,\lambda_n}$ to be asymptotically loss efficient is that (3.11) does not hold for any α_{n1} and α_{n2}. Of course, (3.11) is almost impossible to check. In the following theorem we provide some sufficient conditions for the asymptotic loss efficiency of the GIC_{λ_n}.

Theorem 2. *Suppose that* (2.6) *and* (3.10) *hold and that* $\hat{\sigma}_n^2 \not\to_p 0$ *and* $\hat{\sigma}_n^2 \not\to_p \infty$.
(i) *A sufficient condition for the asymptotic loss efficiency of* $\hat{\alpha}_{n,\lambda_n}$ *is that* (2.5) *holds and* λ_n *is chosen to satisfy*

$$\lambda_n \to \infty \quad \textit{and} \quad \frac{\lambda_n p_n}{n} \to 0. \qquad (3.12)$$

(ii) *If* $\mathcal{A}_n$ *contains at least one correct model with fixed dimension for sufficiently large* n, $\lambda_n \to \infty$ *and* $\lambda_n/n \to 0$, *then* $\hat{\alpha}_{n,\lambda_n}$ *is consistent.*

Remark 3. Unlike the case of $\lambda_n \equiv 2$, it is not required in Theorem 2 that $\hat{\sigma}_n^2$ be a consistent estimator of σ^2.

We now apply Theorems 1 and 2 to Examples 1-3.

Example 1. (continued) We use the notation given by (1.1). In this example (2.4), (2.5), (2.6) and (3.10) hold. Note that $\mathcal{A}_n^c$ is not empty and consistency in the sense of (2.1) is the same as asymptotic loss efficiency in the sense of (2.3) (Proposition 1). By Theorem 1, $\hat{\alpha}_{n,2}$ is consistent if and only if $\bar{\alpha} = \{1, \ldots, p\}$ is the only correct model. By Theorem 2(ii), $\hat{\alpha}_{n,\lambda_n}$ is always consistent if $\lambda_n \to \infty$ and $\lambda_n/n \to 0$.

Example 2. (continued) Note that $n = kr \to \infty$ means that either $k \to \infty$ or $r \to \infty$. Using Theorems 1 and 2, we now show that $\hat{\alpha}_{n,2}$ is better when $k \to \infty$, whereas $\hat{\alpha}_{n,\lambda_n}$ with λ_n satisfying (3.12) is better when $r \to \infty$.

It is easy to see that (2.6) and (3.10) hold. Condition (2.5) holds if k is fixed. If $k \to \infty$, then (2.5) is the same as

$$\liminf_k \frac{1}{k}\sum_{j=1}^{k}(\mu_j - \frac{1}{k}\sum_{i=1}^{k}\mu_i)^2 > 0,$$

which is a reasonable condition.

Consider first the case where $k \to \infty$ and r is fixed. Since the difference in dimensions of the two models in $\mathcal{A}_n$ is $k-1$, an application of Theorem 1(i)&(iv) shows that $\hat{\alpha}_{n,2}$ is always asymptotically loss efficient. On the other hand, it can be shown that if $\lambda_n \to \infty$, then $P\{\hat{\alpha}_{n,\lambda_n} = \alpha_1\} \to 1$. Hence $\hat{\alpha}_{n,\lambda_n}$ is asymptotically loss efficient if and only if the one-mean model is correct.

Next, consider the case where $r \to \infty$ and k is fixed. In this case the dimensions of both models are fixed. By Proposition 1, consistency is the same as asymptotic loss efficiency. By Theorem 2, $\hat{\alpha}_{n,\lambda_n}$ with $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ is consistent. By Theorem 1, $\hat{\alpha}_{n,2}$ is consistent if and only if the one-mean model is incorrect.

Finally, consider the case where $k \to \infty$ and $r \to \infty$. Since $p_n/n = r^{-1} \to 0$, consistency is the same as asymptotic loss efficiency (Proposition 1). By Theorems 1 and 2, both $\hat{\alpha}_{n,2}$ and $\hat{\alpha}_{n,\lambda_n}$ are consistent, but λ_n has to be chosen so that (3.12) holds, i.e., $\lambda_n/r \to 0$. For example, if we choose $\lambda_n = \log n$ (GIC$_{\lambda_n}$ is then equivalent to the BIC in Schwarz (1978)), then $\hat{\alpha}_{n,\lambda_n}$ is inconsistent if $\log n/r \not\to 0$. This is exactly what was described in Section 3 of Stone (1979).

Example 3. (continued) In this case $p_n \to \infty$ as $n \to \infty$. Conditions (2.6) and (3.10) are usually satisfied with $m = 2$. If there exists a correct model in $\mathcal{A}_n$ for some n, then there are many correct models in $\mathcal{A}_n$ and by Theorems 1 and 2, $\hat{\alpha}_{n,\lambda_n}$ is consistent but $\hat{\alpha}_{n,2}$ is not. On the other hand, if there is no correct model in $\mathcal{A}_n$ for all n, then $\hat{\alpha}_{n,2}$ is asymptotically loss efficient but $\hat{\alpha}_{n,\lambda_n}$ may not, since condition (2.5) may not hold.

In conclusion, the GIC$_{\lambda_n}$ method with $\lambda_n \equiv 2$ is more useful in the case where there is no fixed-dimension correct model, whereas the GIC$_{\lambda_n}$ method with $\lambda_n \to \infty$ is more useful in the case where there exist fixed-dimension correct models.

To end this section, we discuss briefly the GIC$_{\lambda_n}$ with $\lambda_n \equiv \lambda$, a constant larger than 2. It is apparent that the GIC$_\lambda$ with a fixed $\lambda > 2$ is a compromise between the GIC$_2$ and the GIC$_{\lambda_n}$ with $\lambda_n \to \infty$. The asymptotic behavior of the GIC$_\lambda$, however, is not as good as the GIC$_2$ in the case where no fixed-dimension correct model exists, and not as good as the GIC$_{\lambda_n}$ when there are fixed-dimension correct models. This can be seen from the proofs of Theorems 1 and 2 in Section 6.

4. Other Selection Methods

In this section we show that some selection methods cited in Section 1 have the same asymptotic behavior (in terms of consistency and asymptotic loss efficiency) as the GIC$_{\lambda_n}$ under certain conditions.

First, consider the GIC$_{\lambda_n}$ with the following particular choice of $\hat{\sigma}_n^2$:

$$\hat{\sigma}_n^2 = \frac{S_n(\bar{\alpha}_n)}{n - p_n}, \tag{4.1}$$

where $\bar{\alpha}_n = \{1, \ldots, p_n\}$. If (4.1) is used, then the GIC$_2$ is the C$_p$ method (Mallows (1973)) and the GIC$_{\lambda_n}$ is the GIC in Rao and Wu (1989). The estimator in (4.1), however, is not necessarily consistent for σ^2 if $\bar{\alpha}_n$ is an incorrect model. Asymptotic behavior of the C$_p$ ($\lambda_n \equiv 2$) is given in the following result.

Theorem 1A. (i) *If* $\Delta_n(\bar{\alpha}_n) \to 0$ *and* $p_n/n \not\to 1$, *then* $\hat{\sigma}_n^2$ *in* (4.1) *is consistent for* σ^2 *and, therefore, the assertions* (i)-(iv) *in Theorem* 1 *are valid for the* C_p. (ii) *If* (2.4) *holds, then the assertions* (i)-(iv) *in Theorem* 1 *are valid for the* C_p.

Note that in Theorem 2, we do not need $\hat{\sigma}_n^2$ to be consistent. Hence we have the following result for the case where $\lambda_n \to \infty$.

Theorem 2A. *Assume that* (2.6) *and* (3.10) *hold. Then the assertions* (i)-(ii) *in Theorem* 2 *are valid for the GIC*$_{\lambda_n}$ *with* $\hat{\sigma}_n^2$ *given by* (4.1) *and* $\lambda_n \to \infty$.

If we use

$$\hat{\sigma}_n^2 = \hat{\sigma}_n^2(\alpha) = \frac{S_n(\alpha)}{n - p_n(\alpha)}$$

(an estimate of σ^2 depends on the model α) in (3.1), then we select a model by minimizing

$$\tilde{\Gamma}_{n,\lambda_n}(\alpha) = \frac{S_n(\alpha)}{n}\left[1 + \frac{\lambda_n p_n(\alpha)}{n - p_n(\alpha)}\right].$$

If $\lambda_n p_n/n \to 0$, this method has the same asymptotic behavior as the method minimizing

$$\log\frac{S_n(\alpha)}{n} + \frac{\lambda_n p_n(\alpha)}{n - p_n(\alpha)},$$

since $\log(1+x) \approx x$ as $x \to 0$. Minimizing $\tilde{\Gamma}_{n,\lambda_n}(\alpha)$ is known as the AIC if $\lambda_n \equiv 2$ and the BIC if $\lambda_n = \log n$.

Let $\tilde{\alpha}_{n,\lambda_n}$ be the model selected by minimizing $\tilde{\Gamma}_{n,\lambda_n}(\alpha)$ over $\alpha \in \mathcal{A}_n$. We have the following result similar to Theorems 1 and 2.

Theorem 3. *Suppose that* (2.6) *holds.*
(i) *The assertions* (i)-(iv) *in Theorem* 1 *are valid for* $\tilde{\alpha}_{n,2}$ (*the AIC*) *if either* (2.4) *holds or*

$$\max_{\alpha\in\mathcal{A}_n}\Delta_n(\alpha)\to 0 \quad \textit{and} \quad \frac{p_n}{n}\not\to 1. \tag{4.2}$$

(ii) *Assume that* (3.10) *holds. Then the assertions* (i)-(ii) *in Theorem* 2 *are valid for* $\tilde{\alpha}_{n,\lambda_n}$ *with* $\lambda_n\to\infty$.

The delete-1 CV method selects a model by minimizing

$$CV_{n,1}(\alpha)=\frac{\|[\boldsymbol{I}_n-\tilde{\boldsymbol{H}}_n(\alpha)]^{-1}[\boldsymbol{y}_n-\hat{\boldsymbol{\mu}}_n(\alpha)]\|^2}{n}$$

over $\alpha\in\mathcal{A}_n$, where $\tilde{\boldsymbol{H}}_n(\alpha)$ is a diagonal matrix whose ith diagonal element is the ith diagonal element of $\boldsymbol{H}_n(\alpha)$. The GCV method replaces $\tilde{\boldsymbol{H}}_n(\alpha)$ by $[n^{-1}\mathrm{tr}\tilde{\boldsymbol{H}}(\alpha)]\boldsymbol{I}_n=[n^{-1}p_n(\alpha)]\boldsymbol{I}_n$, where tr$A$ is the trace of the matrix A, and hence it selects a model by minimizing

$$GCV_n(\alpha)=\frac{S_n(\alpha)}{n[1-n^{-1}p_n(\alpha)]^2}.$$

From the identity

$$\frac{1}{[1-n^{-1}p_n(\alpha)]^2}=1+\frac{2p_n(\alpha)}{n-p_n(\alpha)}+\left[\frac{p_n(\alpha)}{n-p_n(\alpha)}\right]^2,$$

we know that the GCV and the AIC have the same asymptotic behavior if

$$\max_{\alpha\in\mathcal{A}_n}\left[\frac{p_n(\alpha)}{n-p_n(\alpha)}\right]^2\Big/\left[1+\frac{2p_n(\alpha)}{n-p_n(\alpha)}\right]\to 0, \tag{4.3}$$

which holds if and only if (2.4) holds.

Theorem 4. *Suppose that* (2.6) *holds.*
(i) *The assertions* (i)-(iv) *in Theorem* 1 *are valid for the GCV if either* (2.4) *or* (4.2) *holds.*
(ii) *Assume that*

$$h_n=\max_{i\le n}\boldsymbol{x}_i'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_i\to 0. \tag{4.4}$$

Then the assertions (i)-(iv) *in Theorem* 1 *are valid for the delete*-1 *CV.*

Condition (4.4) is stronger than condition (2.4). When neither (2.4) nor (4.2) holds, the GCV and the delete-1 CV may not be asymptotically loss efficient.

Example 2. (continued) We consider Example 2 in the situation where k is large but r, the number of replication, is fixed. Since $p_n/n=r^{-1}$, (2.4) does not hold.

Assume that $\lim_{n\to\infty}\Delta_n(\alpha_1)=\Delta>0$, i.e., (4.2) does not hold. Let y_{ij} be the jth observation in the ith group, $j=1,\ldots,r$, $i=1,\ldots,k$, $\bar{y}_i$ be the ith group mean, $\bar{y}$ be the overall mean, $SS_1=\sum_{i=1}^k\sum_{j=1}^r(y_{ij}-\bar{y})^2$ and $SS_k=\sum_{i=1}^k\sum_{j=1}^r(y_{ij}-\bar{y}_i)^2$. The delete-1 CV and the GCV are identical in this case and select the one-mean model if and only if $SS_1/(1-n^{-1})^2<SS_k/(1-r^{-1})^2$. From

$$\frac{L_n(\alpha_1)}{L_n(\alpha_k)}\to_p\frac{r\Delta}{\sigma^2},\qquad\frac{SS_1}{n}\to_p\sigma^2+\Delta\quad\text{and}\quad\frac{SS_k}{n}\to_p\frac{(r-1)\sigma^2}{r},$$

the delete-1 CV (or the GCV) is not asymptotically loss efficient if $\sigma^2/r<\Delta\leq\sigma^2/(r-1)$.

The delete-d CV is an extension of the delete-1 CV. Suppose that we split the $n\times(1+p_n)$ matrix $(\boldsymbol{y}_n,\boldsymbol{X}_n)$ into two distinct sub-matrices: a $d\times(1+p_n)$ matrix $(\boldsymbol{y}_{n,s},\boldsymbol{X}_{n,s})$ containing the rows of $(\boldsymbol{y}_n,\boldsymbol{X}_n)$ indexed by the integers in s, a subset of $\{1,\ldots,n\}$ of size d, and an $(n-d)\times(1+p_n)$ matrix $(\boldsymbol{y}_{n,s^c},\boldsymbol{X}_{n,s^c})$ containing the rows of $(\boldsymbol{y}_n,\boldsymbol{X}_n)$ indexed by the integers in s^c, the complement of s. For any $\alpha\in\mathcal{A}_n$, we estimate $\boldsymbol{\beta}_n(\alpha)$ by $\hat{\boldsymbol{\beta}}_{n,s^c}(\alpha)$, the LSE based on $(\boldsymbol{y}_{n,s^c},\boldsymbol{X}_{n,s^c})$ under model α. The model is then assessed by $\|\boldsymbol{y}_{n,s}-\hat{\boldsymbol{\mu}}_{n,s}(\alpha)\|^2$, where $\hat{\boldsymbol{\mu}}_{n,s}(\alpha)=\boldsymbol{X}_{n,s}(\alpha)\hat{\boldsymbol{\beta}}_{n,s^c}(\alpha)$ and $\boldsymbol{X}_{n,s}(\alpha)$ is a $d\times p_n(\alpha)$ matrix containing the columns of $\boldsymbol{X}_{n,s}$ indexed by the integers in α. Let $\mathcal{S}$ be a class of N subsets s. The delete-d CV method selects a model by minimizing

$$CV_{n,d}(\alpha)=\frac{1}{dN}\sum_{s\in\mathcal{S}}\|\boldsymbol{y}_{n,s}-\hat{\boldsymbol{\mu}}_{n,s}(\alpha)\|^2$$

over $\alpha\in\mathcal{A}_n$. The set $\mathcal{S}$ can be obtained by using a balanced incomplete block design (Shao (1993)) or by taking a simple random sample from the collection of all possible subsets of $\{1,\ldots,n\}$ of size d (Burman (1989), Shao (1993)).

While the delete-1 CV has the same asymptotic behavior as the C_p (Theorem 4), the delete-d CV has the same asymptotic behavior as the GIC_{λ_n} with

$$\lambda_n=\frac{n}{n-d}+1. \tag{4.5}$$

If $d/n\to0$, then $\lambda_n\to2$; if $d/n\to\tau\in(0,1)$, then $\lambda_n\to\frac{1}{1-\tau}+1$, a fixed constant larger than 2; if $d/n\to1$, then $\lambda_n\to\infty$.

In view of the discussion (in the end of Section 3) for the GIC_λ with a fixed $\lambda>2$, we consider only the case where d is chosen so that $d/n\to1$ $(\lambda_n\to\infty)$.

Theorem 5. *Suppose that* (2.5), (2.6) *and* (3.10) *hold and that*

$$\max_{s\in\mathcal{S}}\sup_{\|c\|=1}\left|\frac{\|\boldsymbol{X}_{n,s}c\|^2}{d}-\frac{\|\boldsymbol{X}_{n,s^c}c\|^2}{n-d}\right|\to0.$$

Then the delete-d CV is asymptotically loss efficient if d is chosen so that

$$\frac{d}{n} \to 1 \quad \textit{and} \quad \frac{p_n}{n-d} \to 0. \tag{4.6}$$

If, in addition, $\mathcal{A}_n$ contains at least one correct model with fixed dimension, then the delete-d CV is consistent.

Remark 4. Condition (4.6) implies condition (2.4) and is similar to condition (3.12) in Theorem 2. In fact, $p_n/(n-d) \to 0$ is a very natural requirement for using the delete-d CV, since $n-d$ is the number of observations used to fit an initial model with as many as p_n parameters.

The PMDL and PLS methods (Rissanen (1986), Wei (1992)) are shown to have the same asymptotic behavior as the BIC method (which is a special case of the GIC) under some situations (Wei (1992)). However, these two methods are intended for the case where $\boldsymbol{e}_n$ is a time series so that the observations have a natural order. Hence, we do not discuss these methods here.

In conclusion, the methods discussed so far can be classified into the following three classes according to their asymptotic behaviors:

Class 1. The GIC$_2$, the C$_p$, the AIC, the delete-1 CV, and the GCV.

Class 2. The GIC$_{\lambda_n}$ with $\lambda_n \to \infty$ and the delete-d CV with $d/n \to 1$.

Class 3. The GIC$_\lambda$ with a fixed $\lambda > 2$ and the delete-d CV with $d/n \to \tau \in (0,1)$.

The methods in class 1 are useful in the case where there is no fixed-dimension correct model. With a suitable choice of λ_n or d, the methods in class 2 are useful in the case where there exist fixed-dimension correct models. The methods in class 3 are compromises of the methods in class 1 and the methods in class 2; but their asymptotic performances are not as good as those of the methods in class 1 in the case where no fixed-dimension correct model exists, and not as good as those of the methods in class 2 when there are fixed-dimension correct models.

5. Empirical Results

We study the magnitude of $P\{\hat{\alpha}_n = \alpha_n^L\}$ with a fixed n by simulation in two examples. Although some selection methods are shown to have the same asymptotic behavior, their fixed sample performances (in terms of $P\{\hat{\alpha}_n = \alpha_n^L\}$) may be different.

The first example is the linear regression (Example 1) with $p = 5$; that is,

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i, \quad i = 1, \ldots, 40,$$

where e_i are independent and identically distributed as $N(0,1)$, x_{ij} is the ith value of the jth explanatory variable x_j, $x_{i1} \equiv 1$, and the values of x_{ij}, $j = 2,3,4,5$, are taken from an example in Gunst and Mason (1980) (also, see Table 1 in Shao (1993)). This study is an extension of that in Shao (1993) which studies the cross-validation methods only.

Table 1. Selection probabilities in regression problem based on 1000 simulations.

True $\boldsymbol{\beta}'$	Model	AIC	C$_p$	GIC	CV$_1$	GCV	CV$_d$
$(2,0,0,4,0)$	1,4**	.567	.594	.804	.484	.587	.934
	1,2,4*	.114	.110	.049	.133	.112	.025
	1,3,4*	.126	.113	.065	.127	.117	.026
	1,4,5*	.101	.095	.057	.138	.097	.012
	1,2,3,4*	.030	.028	.009	.049	.026	.000
	1,2,4,5*	.030	.027	.007	.029	.028	.001
	1,3,4,5*	.022	.026	.008	.030	.026	.002
	1,2,3,4,5*	.010	.007	.001	.009	.007	.000
$(2,0,0,4,8)$	1,4,5**	.683	.690	.881	.641	.691	.947
	1,2,4,5*	.143	.129	.045	.158	.130	.032
	1,3,4,5*	.116	.142	.067	.138	.143	.020
	1,2,3,4,5*	.058	.039	.007	.063	.036	.001
$(2,9,0,4,8)$	1,4,5	.000	.000	.000	.005	.000	.016
	1,2,4,5**	.794	.817	.939	.801	.824	.965
	1,3,4,5*	.000	.000	.000	.005	.000	.002
	1,2,3,4,5*	.206	.183	.061	.189	.176	.017
$(2,9,6,4,8)$	1,2,3,5	.000	.000	.000	.000	.000	.002
	1,2,4,5	.000	.000	.000	.000	.000	.005
	1,3,4,5	.000	.000	.000	.015	.000	.045
	1,2,3,4,5**	1.00	1.00	1.00	.985	1.00	.948

* A correct model

** The optimal correct model

Six selection procedures, the AIC, the C$_p$, the GIC$_{\lambda_n}$ with $\hat{\sigma}_n^2$ given by (4.1), the delete-1 CV (denoted by CV$_1$), the GCV, and the delete-d CV (denoted by CV$_d$), are applied to select a model from $2^p - 1 = 31$ models. The λ_n in the GIC is chosen to be $\log n = \log 40 \approx 3.8$ so that this GIC is almost the same as the BIC. The d in the delete-d CV is chosen to be 25 so that (4.5) approximately holds and the delete-d CV is comparable with the GIC. The $\mathcal{S}$ in the delete-d CV is obtained by taking a random sample of size $2n = 80$ from all possible subsets of $\{1,\ldots,40\}$ of size 25. For these six selection procedures, the empirical probabilities (based on 1,000 simulations) of selecting each model are reported in Table 1, where each model is denoted by a subset of $\{1,\ldots,5\}$ that contains the indices of the explanatory variables x_j in the model. Models corresponding to zero empirical probabilities for all the methods in the simulation are omitted.

The second example considered is the polynomial approximation to a possibly nonlinear curve (Example 3); that is, we select a model from the following class of models:

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_{h-1} x_i^{h-1} + e_i, \quad h = 1, \ldots, p_n. \tag{5.1}$$

In the simulation, $n = 40$ and $p_n = 5$. Other settings and the selection procedures considered are the same as those in the first example. The values of x_i are taken to be the same as x_{i2} in the first example. We consider situations where one of the models in (5.1) is correct, as well as the case where the true model is

$$y_i = \exp(2x_i) + e_i$$

so that none of the models in (5.1) is correct. The results are reported in Table 2.

Table 2. Selection probabilities in polynomial approximation problem based on 1000 simulations.

True $E(y_i)$	Model	AIC	C$_p$	GIC	CV$_1$	GCV	CV$_d$
1	$h = 1$**	.718	.728	.910	.699	.731	.969
	$h = 2$*	.124	.124	.066	.014	.155	.031
	$h = 3$*	.063	.060	.014	.102	.061	.000
	$h = 4$*	.040	.036	.007	.024	.033	.000
	$h = 5$*	.055	.052	.003	.020	.046	.000
$1 + 2x_i$	$h = 2$**	.725	.758	.916	.738	.762	1.00
	$h = 3$*	.124	.117	.060	.170	.118	.000
	$h = 4$*	.084	.070	.015	.065	.069	.000
	$h = 5$*	.067	.055	.009	.027	.051	.000
$1 + 2x_i + 2x_i^2$	$h = 3$**	.742	.758	.917	.763	.760	1.00
	$h = 4$*	.156	.149	.063	.189	.150	.000
	$h = 5$*	.102	.093	.020	.048	.090	.000
$1 + 2x_i + 2x_i^2$	$h = 3$	.000	.000	.000	.000	.000	.006
$+3x_i^3/2$	$h = 4$**	.821	.835	.935	.834	.839	.994
	$h = 5$*	.179	.165	.065	.166	.161	.000
$1 + 2x_i + 2x_i^2$	$h = 4$	.000	.000	.000	.000	.000	.093
$+3x_i^3/2 + 2x_i^4/3$	$h = 5$**	1.00	1.00	1.00	1.00	1.00	.907
$\exp(2x_i)$	$h = 5$	1.00	1.00	1.00	1.00	1.00	1.00

* A correct model

** The optimal correct model

The following is a summary of the results in Tables 1 and 2.

(1) The procedures in class 2 (the GIC and the CV$_d$) have much better empirical performances than the procedures in class 1 (the AIC, the C$_p$, the CV$_1$, and

the GCV) when there are at least two fixed-dimension correct models. The probability $P\{\hat{\alpha}_n = \alpha_n^L\}$ may be very low for the methods in class 1 when the dimension of the optimal model is not close to p_n. This confirms the asymptotic results established in Sections 3 and 4.

(2) The performances of two methods in class 2 may be substantially different. For example, the probability of the GIC selecting the optimal model can be as low as 0.804 in the first example, whereas the CV$_d$ selects the optimal model with probability higher than 0.90 in all cases. On the other hand, the CV$_d$ selects an incorrect model sometimes with a small chance.

6. Proofs

Proof of Proposition 1. We only need to show that (2.3) does not hold, assuming that (2.1) does not hold. If $\mathcal{A}_n^c$ contains exactly one model, then by (2.4), $L_n(\alpha_n^L) \to_p 0$; but by (2.5), $L_n(\hat{\alpha}_n) \not\to_p 0$. Hence (2.3) does not hold. Next, assume that $\mathcal{A}_n^c$ contains more than one models but $p_n(\alpha_n^L) \not\to_p \infty$. Since $P\{\hat{\alpha}_n \neq \alpha_n^L\} \not\to 0$, there exists $\alpha_{n1} \in \mathcal{A}_n^c$ such that $\alpha_{n1} \neq \alpha_n^L$ and $P\{\hat{\alpha}_n = \alpha_{n1}\} \not\to 0$. Then

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^L)} - 1 \geq \left[\frac{L_n(\alpha_{n1})}{L_n(\alpha_n^L)} - 1\right] I(\hat{\alpha}_n = \alpha_{n1}) = \left[\frac{\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha_{n1})\boldsymbol{e}_n}{\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha_n^L)\boldsymbol{e}_n} - 1\right] I(\hat{\alpha}_n = \alpha_{n1}) \not\to_p 0.$$

Proof of (3.2). Note that

$$\Gamma_{n,2}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + \frac{2(\hat{\sigma}_n^2 - \sigma^2)p_n(\alpha)}{n} + \frac{2[\sigma^2 p_n(\alpha) - \boldsymbol{e}_n' \boldsymbol{H}_n(\alpha)\boldsymbol{e}_n]}{n} + \frac{2\boldsymbol{e}_n'[\boldsymbol{I}_n - \boldsymbol{H}_n(\alpha)]\boldsymbol{\mu}_n}{n}.$$

Hence (3.2) follows from

$$\max_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{|\hat{\sigma}_n^2 - \sigma^2| p_n(\alpha)}{n L_n(\alpha)} \to_p 0, \tag{6.1}$$

$$\max_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{|\sigma^2 p_n(\alpha) - \boldsymbol{e}_n' \boldsymbol{H}_n(\alpha)\boldsymbol{e}_n|}{n L_n(\alpha)} \to_p 0, \tag{6.2}$$

and

$$\max_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{|\boldsymbol{e}_n'[\boldsymbol{I}_n - \boldsymbol{H}_n(\alpha)]\boldsymbol{\mu}_n|}{n L_n(\alpha)} \to_p 0. \tag{6.3}$$

Result (6.1) follows from (6.2), $\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha)\boldsymbol{e}_n \leq n L_n(\alpha)$, and the fact that $\hat{\sigma}_n^2 - \sigma^2 \to_p 0$. Results (6.2) and (6.3) can be shown using the same argument in Li (1987), p.970 under condition (2.6).

Proof of Theorem 1. The first statement in (i) is proved in Section 3. The second statement in (i) is a consequence of the first statement and Proposition 1.

For (ii), it suffices to show that

$$\Gamma_{n,2}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + o_p(L_n(\alpha))$$

uniformly in $\alpha \in \mathcal{A}_n^c$, which follows from either

$$\max_{\alpha\in\mathcal{A}_n^c}\left|\frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{p_n(\alpha)} - \sigma^2\right| \to_p 0 \qquad (6.4)$$

or

$$\max_{\alpha\in\mathcal{A}_n^c,\alpha\neq\alpha_n^c}\left|\frac{\boldsymbol{e}_n'[\boldsymbol{H}_n(\alpha) - \boldsymbol{H}_n(\alpha_n^c)]\boldsymbol{e}_n}{p_n(\alpha) - p_n(\alpha_n^c)} - \sigma^2\right| \to_p 0. \qquad (6.5)$$

From Theorem 2 of Whittle (1960),

$$E\left|\frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{p_n(\alpha)} - \sigma^2\right|^{2m} \leq \frac{c}{[p_n(\alpha)]^m}, \qquad (6.6)$$

where c is a positive constant. Then for any $\epsilon > 0$,

$$P\left\{\max_{\alpha\in\mathcal{A}_n^c}\left|\frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{p_n(\alpha)} - \sigma^2\right| > \epsilon\right\} \leq c\epsilon^{-2m}\sum_{\alpha\in\mathcal{A}_n^c}\frac{1}{[p_n(\alpha)]^m}.$$

Hence (6.4) is implied by condition (3.4). A similar argument shows that (6.5) is implied by condition (3.5).

The result in (iii) can be proved using the same argument in Example 4. For (iv), it suffices to show that $p_n(\alpha_n^c) \to \infty$ is the same as (3.4) and $\min_{\alpha\in\mathcal{A}_n^c,\alpha\neq\alpha_n^c} p_n(\alpha) - p_n(\alpha_n^c) \to \infty$ is the same as (3.5), which is apparent if the number of models in $\mathcal{A}_n^c$ is bounded. The proof for the case where $m = 2$ and $\mathcal{A}_n = \{\alpha_i, i = 1,\ldots,p_n\}$ with $\alpha_i = \{1,\ldots,i\}$ is the same as that in Li (1987), p.963.

Proof of Theorem 2. From (6.6) and condition (3.10),

$$\frac{\boldsymbol{e}_n'\boldsymbol{H}_n(\alpha)\boldsymbol{e}_n}{\lambda_n p_n(\alpha)} = O_p(\lambda_n^{-1})$$

uniformly in $\alpha \in \mathcal{A}_n^c$. Hence (3.9) holds. Since $L_n(\alpha) > \Delta_n(\alpha)$, (3.7) and conditions (2.5) and (3.12) imply that $\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha) + o_p(L_n(\alpha))$ uniformly in $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$, and if $\mathcal{A}_n^c$ is not empty, $\Gamma_{n,\lambda_n}(\alpha_n^c) = o_p(L_n(\alpha))$ uniformly in $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$. The result in (i) is established.

If $\mathcal{A}_n$ contains at least one correct model with fixed dimension, then (2.5) holds and $P\{\hat\alpha_{n,\lambda_n} = \alpha_n^c\} \to 1$. The consistency of $\hat\alpha_{n,\lambda_n}$ follows from the fact that $\alpha_n^L = \alpha_n^c$.

Proofs of Theorems 1A, 2A, and 3. First, consider Theorem 1A. Note that

$$\frac{S_n(\bar{\alpha}_n)}{n-p_n} = \frac{\boldsymbol{e}_n'[\boldsymbol{I}_n - \boldsymbol{H}_n(\bar{\alpha}_n)]\boldsymbol{e}_n}{n-p_n} + \frac{n}{n-p_n}\Delta_n(\bar{\alpha}_n) + \frac{2\boldsymbol{e}_n'[\boldsymbol{I}_n - \boldsymbol{H}_n(\bar{\alpha}_n)]\boldsymbol{\mu}_n}{n-p_n}. \quad (6.7)$$

If $\Delta_n(\bar{\alpha}_n) \to 0$ and $p_n/n \not\to 1$, then $S_n(\bar{\alpha}_n)/(n-p_n)$ is consistent and the result in (i) follows.

Suppose now that (2.4) holds. For the result in (ii), it suffices to show that (6.1) still holds for $\hat{\sigma}_n^2 = S_n(\bar{\alpha}_n)/(n-p_n)$. From (6.7), $S_n(\bar{\alpha}_n)/(n-p_n) = \sigma^2 + o_p(1) + O_p(\Delta_n(\bar{\alpha}_n))$. Hence (6.1) follows from the fact that

$$\max_{\alpha\in\mathcal{A}_n} \frac{\Delta_n(\bar{\alpha}_n)p_n(\alpha)}{nL_n(\alpha)} \leq \frac{p_n}{n}.$$

The proofs for Theorem 2A and Theorem 3 are similar.

Proof of Theorem 4. (i) If (2.4) holds, then (4.3) holds and the result follows from Theorem 3(i). Now, assume that (4.2) holds. Then $p_n(\alpha_n^L)/n \to_p 0$ and $p_n(\hat{\alpha}_n)/n \to_p 0$, where $\hat{\alpha}_n$ is the model selected by the GCV. If $\mathcal{A}_n^c$ is empty for all n, then

$$GCV_n(\hat{\alpha}_n) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\hat{\alpha}_n) + o_p(L_n(\hat{\alpha}_n)),$$

$$GCV_n(\alpha_n^L) = \frac{\|\boldsymbol{e}_n\|^2}{n} + L_n(\alpha_n^L) + o_p(L_n(\alpha_n^L)),$$

and

$$0 \leq \frac{GCV_n(\hat{\alpha}_n) - GCV_n(\alpha_n^L)}{L_n(\hat{\alpha}_n)} = \frac{L_n(\hat{\alpha}_n) - L_n(\alpha_n^L)}{L_n(\hat{\alpha}_n)} + o_p(1) \leq o_p(1).$$

This proves that (2.3) holds. The proof for the case where $\mathcal{A}_n^c$ is nonempty is similar to the proof of Theorem 1.

(ii) Define

$$T_n(\alpha) = [\boldsymbol{y}_n - \hat{\boldsymbol{\mu}}_n(\alpha)]'\tilde{\boldsymbol{H}}_n(\alpha)[\boldsymbol{y}_n - \hat{\boldsymbol{\mu}}_n(\alpha)].$$

Then

$$CV_{n,1}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{2T_n(\alpha)}{n} + O_p\left(\frac{h_nT_n(\alpha)}{n}\right).$$

The result follows if

$$\frac{T_n(\alpha) - \sigma^2 p_n(\alpha)}{nL_n(\alpha)} = o_p(1) \quad \text{uniformly in } \alpha \in \mathcal{A}_n - \mathcal{A}_n^c, \quad (6.8)$$

$$E\left|\frac{T_n(\alpha)}{p_n(\alpha)} - \sigma^2\right|^{2m} \leq \frac{c}{[p_n(\alpha)]^m} \quad \alpha \in \mathcal{A}_n^c \quad (6.9)$$

and

$$E\left|\frac{T_n(\alpha)-T_n(\alpha_n^c)}{p_n(\alpha)-p_n(\alpha_n^c)}-\sigma^2\right|^{2m}\leq\frac{c}{[p_n(\alpha)-p_n(\alpha_n^c)]^m}\quad\alpha\in\mathcal{A}_n^c \tag{6.10}$$

for some $c>0$ and positive integer m such that $E(y_1-\mu_1)^{4m}<\infty$. Let

$$\boldsymbol{W}_n(\alpha)=[\boldsymbol{I}_n-\boldsymbol{H}_n(\alpha)]\tilde{\boldsymbol{H}}_n(\alpha)[\boldsymbol{I}_n-\boldsymbol{H}_n(\alpha)].$$

When $\alpha\in\mathcal{A}_n-\mathcal{A}_n^c$,

$$T_n(\alpha)=\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{e}_n+2\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{\mu}_n+\boldsymbol{\mu}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{\mu}_n. \tag{6.11}$$

From Theorem 2 of Whittle (1960),

$$E\left|\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{e}_n-E[\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{e}_n]\right|^{2l}\leq c[\mathrm{tr}\boldsymbol{W}_n^2(\alpha)]^l\leq ch_n^l[\mathrm{tr}\boldsymbol{W}_n(\alpha)]^l. \tag{6.12}$$

Note that

$$\mathrm{tr}\boldsymbol{W}_n(\alpha)=\mathrm{tr}\tilde{\boldsymbol{H}}_n(\alpha)[\boldsymbol{I}_n-\boldsymbol{H}_n(\alpha)]=p_n(\alpha)-\mathrm{tr}\tilde{\boldsymbol{H}}_n^2(\alpha)\leq p_n(\alpha). \tag{6.13}$$

By (2.6), (6.12) and (6.13),

$$P\left\{\max_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\left|\frac{\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{e}_n-E[\boldsymbol{e}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{e}_n]}{nR_n(\alpha)}\right|>\epsilon\right\}\leq c\epsilon^{-2l}\sum_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\frac{1}{[nR_n(\alpha)]^l}\to 0.$$

Then (6.8) follows from (6.11), $\boldsymbol{\mu}_n'\boldsymbol{W}_n(\alpha)\boldsymbol{\mu}_n\leq h_n\Delta_n(\alpha)\leq h_nR_n(\alpha)$ and the fact that (2.6) implies

$$\max_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\left|\frac{L_n(\alpha)}{R_n(\alpha)}-1\right|\to_p 0.$$

Results (6.9) and (6.10) follow from Theorem 2 of Whittle (1960), the identity (6.13), and the fact that

$$\mathrm{tr}\tilde{\boldsymbol{H}}_n^2(\alpha)\leq h_np_n(\alpha)\quad\text{and}\quad\mathrm{tr}[\tilde{\boldsymbol{H}}_n^2(\alpha)-\tilde{\boldsymbol{H}}_n^2(\alpha_n^c)]\leq 2h_n[p_n(\alpha)-p_n(\alpha_n^c)]$$

when $\alpha\in\mathcal{A}_n^c$.

Proof of Theorem 5. It follows from the proof in Shao (1993), Appendix that under the given conditions,

$$CV_{n,d}(\alpha)=\frac{S_n(\alpha)}{n}+\frac{\lambda_nT_n(\alpha)}{n}+o_p\left(\frac{\lambda_nT_n(\alpha)}{n}\right)$$

uniformly in $\alpha\in\mathcal{A}_n$, where λ_n is given by (4.5) and $T_n(\alpha)$ is given by (6.11). Then the result follows from the given conditions and result (6.9).

Acknowledgements

I would like to thank the referees and an Associate Editor for their constructive comments and suggestions. The research was supported by National Sciences Foundation Grant DMS-9504425 and National Security Agency Grant MDA904-96-1-0066.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-127.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503-514.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70**, 320-328.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application.* Marcel Dekker, New York.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41**, 190-195.

Li, K.-C. (1987). Asymptotic optimality for C$_p$, C$_L$, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958-975.

Mallows, C. L. (1973). Some comments on C$_p$. *Technometrics* **15**, 661-675.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-765.

Pötscher, B. M. (1989). Model selection under nonstationary: autoregressive models and stochastic linear regression models. *Ann. Statist.* **17**, 1257-1274.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-374.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080-1100.

Schwarz, G. (1978). Estimating the dimensions of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-49.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc. Ser. B* **41**, 276-278.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5**, 302-305.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299-313.

Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison, WI 53706, U.S.A.

(Received March 1994; accepted September 1996)

COMMENT

Rudolf Beran

University of California, Berkeley

Professor Shao's welcome asymptotic analysis of standard model selection procedures divides these into three categories: those that perform better when one or more correct models have fixed dimension under the asymptotics; those that do better when no correct model has fixed dimension; and intermediate methods. Adopting the premise that model selection is intended to reduce estimation risk under quadratic loss, my discussion will draw attention to two points:

- GIC_{λ_n} selection estimators with $\lim_{n\to\infty}\lambda_n = \infty$ can have arbitrarily high asymptotic risk when the signal-to-noise ratio is large enough.
- GIC_{λ_n} selection estimators with either $\lambda_n = 2$ or $\lim_{n\to\infty}\lambda_n = \infty$ are not asymptotically minimax unless the signal to noise ratio converges to zero. They are dominated, in maximum risk, by a variety of procedures that taper the components of the least squares fit toward zero.

I will develop both points in a signal recovery setting that is formally a special case of Shao's problem. Suppose that $X_n = \{X_{n,t} : t \in T_n\}$ is an observation on a discrete signal $\xi_n = \{\xi_{n,t} : t \in T_n\}$ that is measured with error at the time points $T_n = \{1, \dots, n\}$. The measurement errors are independent and are such that the distribution of each component $X_{n,t}$ is $N(\xi_{n,t}, \sigma^2)$.

For any real-valued function f defined on T_n, let $\mathrm{ave}(f) = n^{-1}\sum_{t\in T_n} f(t)$. The time-averaged quadratic loss of any estimator $\hat{\xi}_n$ is then

$$L_n(\hat{\xi}_n, \xi_n) = \mathrm{ave}[(\hat{\xi}_n - \xi_n)^2]$$

and the corresponding risk is

$$R_n(\hat{\xi}_n, \xi_n, \sigma^2) = \mathrm{E}L_n(\hat{\xi}_n, \xi_n).$$

Model selection and related estimators typically have smaller risk when all but a few components of ξ_n are small. With enough prior information, this favorable situation may be approximated by suitable orthogonal transformation of X_n before estimation. This transformation leaves the Gaussian error distribution unchanged. A model selection or other estimator constructed in the new coordinate system may be transformed back to the original coordinate system without changing its quadratic loss. Thus, in signal recovery problems, the $\{X_{n,t}\}$ might be Fourier, or wavelet, or analysis of variance, or orthogonal polynomial coefficients of the observed signal.

Let $u \in [0,1]$. We consider nested model selection, in which the candidate estimators have the form $\hat{\xi}_n(u) = \{\hat{\xi}_{n,t}(u) : t \in T_n\}$, with $\hat{\xi}_{n,t}(u) = X_{n,t}$ whenever $t/(n+1) \leq u$ and $\hat{\xi}_{n,t} = 0$ otherwise. The value of u will be chosen by the GIC_{λ_n} method in Shao's Section 3. Let $\hat{\sigma}_n^2$ be a consistent estimator of σ^2 that satisfies

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2 \leq r} \mathrm{E}|\hat{\sigma}_n^2 - \sigma^2| = 0 \tag{1}$$

for every $r \in [0,\infty)$. Such variance estimators may constructed externally using replication or internally by methods such as those described in Rice (1984). The GIC_{λ_n} selection criterion is

$$\hat{\Gamma}_n(u,\lambda_n) = \hat{\gamma}_n(u) + \lambda_n \hat{\sigma}_n^2 n^{-1}[(n+1)u]_I,$$

where $\hat{\gamma}_n(u) = n^{-1}\sum_{t/(n+1)>u} X_{n,t}^2$ and $[\cdot]_I$ is the integer part function. Let $\hat{u}_n$ be the smallest value of $u \in [0,1]$ that minimizes $\hat{\Gamma}_n(u,\lambda_n)$. Existence of $\hat{u}_n$ is assured because the criterion function assumes only a finite number of values. The model selection estimator $\hat{\xi}_n(\hat{u}_n)$ will be denoted by $\hat{\xi}_{n,\lambda_n}$.

Proposition 1. *In the signal-plus-noise model, with* $\hat{\sigma}_n^2$ *satisfying* (1), *the following bounds hold for every* $r \in [0,\infty)$*:*

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2 \leq r} R_n(\hat{\xi}_{n,2}, \xi_n, \sigma^2) = \sigma^2 \min(r,1). \tag{2}$$

If $\lim_{n\to\infty} \lambda_n = \infty$, *then*

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2 \leq r} R_n(\hat{\xi}_{n,\lambda_n}, \xi_n, \sigma^2) = \sigma^2 r. \tag{3}$$

The least squares estimator X_n *satisfies*

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2 \leq r} R_n(X_n, \xi_n, \sigma^2) = \sigma^2. \tag{4}$$

This proposition will be proved at the end of the discussion. Let us consider some implications:

(a) If ξ_n is a voltage signal, then $\mathrm{ave}(\xi_n^2)$ is the time-averaged power dissipated by this signal in passing through a unit resistance. Consequently, $\mathrm{ave}(\xi_n^2)/\sigma^2$ is the time-averaged signal-to-noise ratio in our signal recovery problem. The maximum risks in Proposition 1 are computed over subsets of ξ_n values that are generated by bounding the signal-to-noise ratio from above.

(b) For $r = 0$, the limiting maximum risks in Proposition 1 do not distinguish between the performance of $\hat{\xi}_{n,2}$ and $\hat{\xi}_{n,\lambda_n}$ with $\lim_{n\to\infty}\lambda_n = \infty$. Theorems 1 and 2 in Shao's paper indicate that the latter estimators may perform better in some (but not all) circumstances where the signal-to-noise ratio converges to zero.

(c) As long as the signal-to-noise ratio does not exceed 1, both $\hat{\xi}_{n,2}$ and $\hat{\xi}_{n,\lambda_n}$ with $\lim_{n\to\infty}\lambda_n = \infty$ have the same asymptotic maximum risk. Once the signal to noise ratio exceeds 1, then $\hat{\xi}_{n,\lambda_n}$ has greater asymptotic maximum risk than $\hat{\xi}_{n,2}$ or even the least squares estimator X_n.

(d) For all values of r, the asymptotic maximum risk of $\hat{\xi}_{n,\lambda_n}$ with $\lim_{n\to\infty}\lambda_n = \infty$ coincides with that of the trivial estimator $\hat{\xi}_n = 0$. This does not mean that $\hat{\xi}_{n,\lambda_n}$ is trivial.

(e) For all values of r, the asymptotic maximum risk of $\hat{\xi}_{n,2}$ equals the smaller of the asymptotic maximum risks of X_n and $\hat{\xi}_{n,\lambda_n}$ with $\lim_{n\to\infty}\lambda_n = \infty$. This argument strongly promotes the use of $\hat{\xi}_{n,2}$ over these two competitors unless one is confident that the special circumstances of remark b hold.

How well do model selection estimators perform within the class of *all* estimators of ξ_n? An answer that complements Proposition 1 is

Proposition 2. *In the signal-plus-noise model, with $\hat{\sigma}_n^2$ satisfying* (1)*, the following equality holds for every $r \in [0,\infty)$:*

$$\lim_{n\to\infty} \inf_{\hat{\xi}_n} \sup_{\text{ave}(\xi_n^2)/\sigma^2 \le r} R_n(\hat{\xi}_n, \xi_n, \sigma^2) = \sigma^2 r/(r+1). \tag{5}$$

This result follows from Pinsker's (1980) general lower bound on risk in signal recovery from Gaussian noise. It may also be derived from ideas in Stein (1956) by considering best orthogonally equivariant estimators in the submodel where $\text{ave}(\xi_n^2)/\sigma^2 = r$. To be asymptotically minimax, an estimator $\hat{\xi}_n$ must satisfy

$$\lim_{n\to\infty} \sup_{\text{ave}(\xi_n^2)/\sigma^2 \le r} R_n(\hat{\xi}_n, \xi_n, \sigma^2) = \sigma^2 r/(r+1).$$

Simplest among asymptotically minimax estimators is the James-Stein (1961) estimator

$$\hat{\xi}_{n,S} = [1 - \hat{\sigma}_n^2/\text{ave}(X_n^2)]^+ X_n,$$

where $[\cdot]^+$ denotes the positive part function and $\hat{\sigma}_n^2$ is an estimator of σ^2 that satisfies (1). For every positive, finite r and σ^2, $\sigma^2 r/(r+1) < \sigma^2 \min(r,1)$. Hence, for large n, the James-Stein estimator dominates, in maximum risk, any of the three estimators discussed in Proposition 1. Figure 1 reveals the extent of this domination.

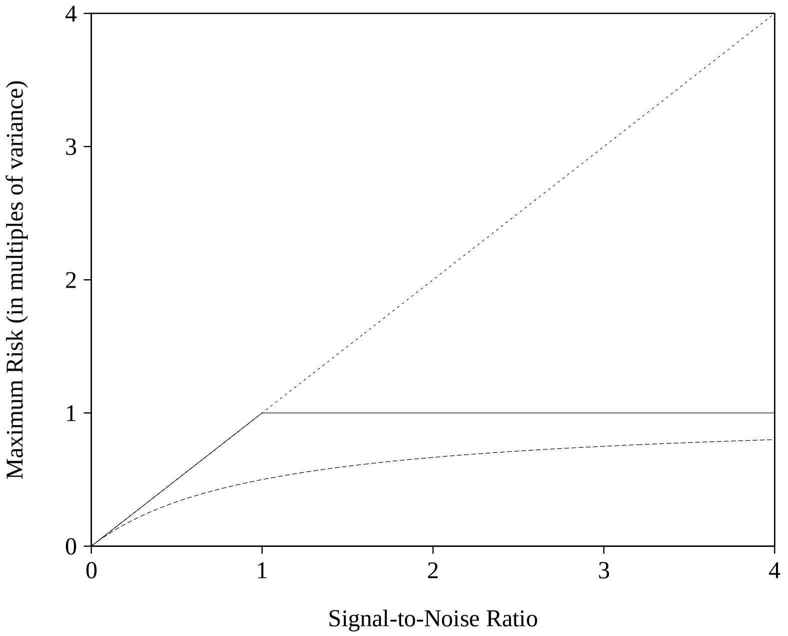

Figure 1. The asymptotic maximum risks of nested model selection estimators generated by the GIC_2 criterion (solid lines) and by the GIC_{λ_n} criterion when $\lim_{n\to\infty}\lambda_n = \infty$ (broken line). The asymptotic minimax risk, attained by good tapered estimators, is the dashed curve below.

The story only begins here. We can construct asymptotically minimax estimators that dominate the James-Stein estimator over submodels. Let $\mathcal{G}_n$ be a given closed convex subset of $[0,1]^{T_n}$ that contains all constants in $[0,1]$. Each function $g \in \mathcal{G}_n$ defines a candidate *modulation* estimator $gX_n = \{g(t)X_{n,t} : t \in T_n\}$ for ξ_n. The risk of this candidate estimator under quadratic loss is

$$R_n(gX_n, \xi_n, \sigma^2) = \mathrm{ave}[\sigma^2 g^2 + \xi_n^2(1-g)^2].$$

Here squaring is done componentwise. An estimator of this risk, suggested by Stein's unbiased estimator for risk or by the C_p idea, is

$$\hat{R}_n(g) = \mathrm{ave}[g^2\hat{\sigma}_n^2 + (1-g)^2(X_n^2 - \hat{\sigma}_n^2)].$$

The proposal is to use the modulation estimator $\hat{g}_n X_n$, where $\hat{g}_n$ minimizes $\hat{R}_n(g)$ over all $g \in \mathcal{G}$. When $\mathcal{G}_n$ consists of all constant functions in $[0,1]^{T_n}$, the modulation estimator $\hat{g}_n X_n$ is just the James-Stein estimator described above.

To improve on James-Stein, let $\mathcal{G}_{n,\mathrm{mon}}$ be the set of all nonincreasing functions in $[0,1]^{T_n}$. The class of candidate estimators $\{gX_n : g \in \mathcal{G}_{n,\mathrm{mon}}\}$ now contains the nested model selection estimators discussed earlier. It contains as well candidate estimators that selectively taper the coordinates of X_n towards zero. Choosing $\hat{g}_{n,\text{mon}}$ to minimize $\hat{R}_n(g)$ over $g \in \mathcal{G}_{n,\text{mon}}$ generalizes precisely choosing $\hat{u}_n$ to minimize $\hat{\Gamma}_n(u, 2)$ over $u \in [0,1]$. Because the class of candidate modulators $\mathcal{G}_{n,\text{mon}}$ contains all constant functions on $[0,1]^{T_n}$, it turns out that $\hat{g}_{n,\text{mon}}X_n$ is asymptotically minimax, unlike $\hat{\xi}_{n,2}$:

$$\lim_{n\to\infty} \sup_{\text{ave}(\xi_n^2)/\sigma^2 \le r} R_n(\hat{g}_{n,\text{mon}}X_n, \xi_n, \sigma^2) = \sigma^2 r/(r+1).$$

Thus, on the one hand, $\hat{g}_{n,\text{mon}}X_n$ dominates, for every $r > 0$, the nested model selection estimators treated in Proposition 1. On the other hand, because $\mathcal{G}_{n,\text{mon}}$ is richer than the class of all constants in $[0,1]^{T_n}$, the maximum risk of the estimator $\hat{g}_{n,\text{mon}}X_n$ asymptotically dominates that of the James-Stein estimator over large classes of submodels within $\text{ave}(\xi_n^2)/\sigma^2 \le r$. For further details on these points, on other interesting choices of $\mathcal{G}_n$, and on algorithms for computing $\hat{g}_n$, see Beran and Dümbgen (1996).

In short, when quadratic risk is the criterion and the signal-to-noise ratio is not asymptotically zero, data-driven tapering of X_n is superior to model selection for estimating ξ_n. This finding is not entirely surprising, since the components of X_n could be Fourier or wavelet coefficients computed from the original data; and tapering is known to reduce the Gibbs phenomenon that is created by truncating a Fourier series.

Proof of Proposition 1. Fix r and suppose throughout that $\text{ave}(\xi_n^2)/\sigma^2 \le r$ for every n. Result (4) is obvious. Let

$$\begin{aligned} V_{n,1}(u) &= n^{-1/2} \sum_{t/(n+1)>u} [(X_{n,t} - \xi_{n,t})^2 - \sigma^2] \\ V_{n,2}(u) &= n^{-1/2} \sum_{t/(n+1)>u} \xi_{n,t}(X_{n,t} - \xi_{n,t}) \end{aligned}$$

for every $0 \le u \le 1$. Let $\|\cdot\|$ denote the supremum norm on $[0,1]$. By Kolmogorov's inequality, there exist finite constants C_1 and C_2 such that $\sup_{\text{ave}(\xi_n^2)/\sigma^2 \le r} \text{E}\|V_{n,i}\| \le C_i$ for $i = 1,2$ and every $n \ge 1$.

First step. Recall the definition of $\hat{\gamma}_n(u)$ and let $\nu_n(u) = n^{-1}\sum_{t/(n+1)>u} \xi_{n,t}^2$. Then

$$\hat{\gamma}_n(u) = \nu_n(u) + \sigma^2(1 - n^{-1}[(n+1)u]_I) + n^{-1/2}V_{n,1}(u) + 2n^{-1/2}V_{n,2}(u). \tag{6}$$

Consequently,

$$\|\hat{\Gamma}_n(\cdot,2) - \nu_n(\cdot) - \sigma^2(1+\cdot)\| = o_E(1). \tag{7}$$

The notation $o_E(1)$ represents a term r_n for which $\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2\le r} \mathrm{E}|r_n| = 0$. Moreover, since

$$L_n(\hat{\xi}_n(u), \xi_n) = \nu_n(u) + \sigma^2 n^{-1}[(n+1)u]_I + n^{-1/2}V_{n,1}(0) - n^{-1/2}V_{n,1}(u), \quad (8)$$

it follows that

$$\|L_n(\hat{\xi}_n(\cdot), \xi_n) + \sigma^2 - \hat{\Gamma}_n(\cdot, 2)\| = o_E(1). \quad (9)$$

On the one hand, from the definition of $\hat{\xi}_{n,2}$, (9), and (7),

$$\begin{aligned} L_n(\hat{\xi}_{n,2}, \xi_n) &= \min_{0\le u\le 1}[\nu_n(u) + \sigma^2 u + \sigma^2] - \sigma^2 + o_E(1) \\ &\le \min(\nu_n(0) + \sigma^2, 2\sigma^2) - \sigma^2 + o_E(1) \\ &\le \sigma^2 \min(r, 1) + o_E(1). \end{aligned}$$

Hence,

$$\limsup_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2\le r} R_n(\hat{\xi}_{n,2}, \xi_n, \sigma^2) \le \sigma^2 \min(r, 1). \quad (10)$$

On the other hand, if $\xi_{n,t}^2 = r\sigma^2$ for every $t \in T_n$, then, starting as in the preceding paragraph,

$$\begin{aligned} L_n(\hat{\xi}_{n,2}, \xi_n) &= \min_{0\le u\le 1}[\nu_n(u) + \sigma^2 u + \sigma^2] - \sigma^2 + o_E(1) \\ &= \sigma^2 \min_{0\le u\le 1}(r + (1-r)u) + o_E(1) \\ &= \sigma^2 \min(r, 1) + o_E(1). \end{aligned}$$

Hence,

$$\limsup_{n\to\infty} \sup_{\mathrm{ave}(\xi_n^2)/\sigma^2\le r} R_n(\hat{\xi}_{n,2}, \xi_n, \sigma^2) \ge \sigma^2 \min(r, 1). \quad (11)$$

Result (2) follows from (10) and (11).

Second step. From (6) and the definition $\hat{\Gamma}_n(u, \lambda_n)$,

$$\|\hat{\Gamma}_n(\cdot, \lambda_n) - \nu_n(\cdot) - \sigma^2(1 - \cdot) - \lambda_n\hat{\sigma}_n^2 n^{-1}[(n+1)\cdot]_I\| = o_E(1). \quad (12)$$

This implies

$$\min_{0\le u\le 1} \hat{\Gamma}_n(u, \lambda_n) - \min_{0\le u\le 1}\{\nu_n(u) + \sigma^2(1-u) + \lambda_n\hat{\sigma}_n^2 n^{-1}[(n+1)u]_I\} = o_E(1),$$

and so

$$\min_{0\le u\le 1} \hat{\Gamma}_n(u, \lambda_n) \le \sigma^2(r+1) + o_E(1).$$

Since $\lambda_n \to \infty$, an argument by contradiction using this bound, (1), and (12) establishes $\hat{u}_n = o_E(1)$. Then, from (8),

$$L_n(\hat{\xi}_{n,\lambda_n}, \xi_n) = \nu_n(\hat{u}_n) + o_E(1). \quad (13)$$

Consequently,

$$\limsup_{n\to\infty} \sup_{\text{ave}(\xi_n^2)/\sigma^2\le r} R_n(\hat{\xi}_{n,\lambda_n}, \xi_n, \sigma^2) \le \sigma^2 r.$$

On the other hand, setting $\xi_{n,n}^2 = nr\sigma^2$ and $\xi_{n,t}^2 = 0$ otherwise in (13) yields

$$\liminf_{n\to\infty} \sup_{\text{ave}(\xi_n^2)/\sigma^2\le r} R_n(\hat{\xi}_{n,\lambda_n}, \xi_n, \sigma^2) \ge \sigma^2 r.$$

Result (3) now follows.

Acknowledgement

This research was supported in part by National Science Foundation Grant DMS95-30492.

Department of Statistics, University of California, Berkeley, CA94720, U.S.A.

COMMENT

J. Sunil Rao and Robert Tibshirani

Cleveland Clinic Foundation and University of Toronto

In an impressive series of papers over the last few years, Jun Shao has shed new light on the behaviour of linear model selection procedures from an asymptotic point of view. This new paper ties together much of this work and in one broad sweep, he develops a framework for comparing the majority of model selection procedures in current use. He is to be congratulated.

Shao's framework is the following. In a linear model setting with parameter vector β, he lets α index possible subsets of β, denoted by $\beta(\alpha)$. He defines a true model α_n^L to be the submodel minimizing the averaged squared prediction error. Then for any model selection procedure producing an estimated subset $\hat{\alpha}_n$, he asks whether the procedure is consistent, that is whether

$$P\{\hat{\alpha}_n = \alpha_n^L\} \to 1. \tag{1}$$

We wonder whether he is asking the right question. Our concern is twofold. First, while consistency seems a reasonable objective we often want our procedure to produce accurate estimates in terms of mean squared (or prediction) error. The two objectives are not equivalent, as overfitting or underfitting can have very different effects on prediction accuracy.

Secondly, rather than focus on a fixed subset, it seems more natural to focus the selection procedure on a complexity parameter. In particular, we construct the cost-complexity criterion

$$C_\lambda(\beta) = RSS(\beta) + \lambda k, \tag{2}$$

where $RSS(\beta)$ is the residual sum of squares for some model β (more precisely for some $\beta(\alpha)$ but notationally supressed for clarity) and k is the number of non-zero elements in β. For a fixed $\lambda > 0$ we can find the parameter β minimizing $C_\lambda(\beta)$. In practice, we use a procedure like cross-validation to find the value of $\hat{\lambda}$ producing the smallest estimated prediction error, and then for our final model we choose the β minimizing $C_{\hat{\lambda}}(\beta)$.

The parameter λ roughly indexes model size. Hence in the above procedure cross-validation is examining the performance of a given *model size*, as opposed to a given *model.* Model size is likely to be a more "portable" quantity than a fixed model, in going between training and validation samples. This cost-complexity approach is the basis of the pruning procedure in the CART work (Breiman et al. (1984)), and was studied in the linear model setting by Rao (1994).

To examine these issues, we reran 100 realizations of the simulation study of Shao's Table 1, producing our own Table 1. We have included in our table the number of underfit and overfit models and the model error $ME = (\mu - \hat{\mu}_n)'(\mu - \hat{\mu}_n)$ where $\mu = \mathbf{X}\beta$ and $\hat{\mu}_n = \mathbf{X}(\hat{\alpha}_n)\hat{\beta}(\hat{\alpha}_n)$. This model error associated with each selection procedure is averaged over the 100 realizations. Along with some of the selection procedures studied by Shao, we have included the full least squares fit, leave-d out cross-validation for a number of different values d, and the adaptive cost complexity parameter (CCP) approach, using leave-out 25 cross-validation to choose λ in the range $[\log n, n/\log n]$. This range is chosen to guarantee consistency of the procedure. The estimators were applied to the four scenarios given by Shao, and a fifth one at the bottom of the table. The results show:

1. Methods with high correct selection probabilities do not always give accurate predictions. In particular, CV$_d$ with large d (25 or 30) sometimes underfits badly, resulting in high model error. The value 25 (chosen by Shao) seems unusually large to us when the sample size is 40: we wonder what Shao's recommendation is for the practitioner?
2. In the last case the full model beats all model selectors. We feel that the full model should be included in any comparison of such procedures.
3. The adaptive CCP method looks to be the overall winner, performing well in terms of both correct model selection and model error.

The area of model selection is very complex, with many aspects not yet (in our view) well understood. Our discussion was meant to raise more questions, in

Table 1. Shao's Example 1 indicating overfitting, underfitting counts, number of times true model selected, and ME for the selection procedure (100 realizations of the simulation).

True Model	Procedure	number	number	number	ME
		overfit	underfit	correct	
$\beta = (2,0,0,4,0)'$	full model	100	0	0	5.231
	C_p	41	0	59	3.943
	BIC$(\lambda = \log n)$	17	0	83	3.031
	GIC$(\lambda = n/\log n)$	1	0	99	2.174
	CV$d(d = 1)$	78	0	22	4.026
	CV$d(d = 20)$	30	0	70	3.113
	CV$d(d = 25)$	22	0	78	2.809
	CV$d(d = 30)$	18	0	82	2.276
	Adaptive CCP	4	0	96	2.386
$\beta = (2,0,0,4,8)'$	full model	100	0	0	5.231
	C_p	28	0	72	4.373
	BIC$(\lambda = \log n)$	10	0	90	3.792
	GIC$(\lambda = n/\log n)$	0	0	100	3.309
	CV$d(d = 1)$	73	0	27	4.427
	CV$d(d = 20)$	23	0	77	3.573
	CV$d(d = 25)$	24	0	76	3.784
	CV$d(d = 30)$	15	1	84	4.002
	Adaptive CCP	4	0	96	3.563
$\beta = (2,9,0,4,8)'$	full model	100	0	0	5.231
	C_p	18	0	82	4.734
	BIC$(\lambda = \log n)$	5	0	95	4.389
	GIC$(\lambda = n/\log n)$	0	1	99	4.412
	CV$d(d = 1)$	39	3	58	5.560
	CV$d(d = 20)$	14	1	85	4.682
	CV$d(d = 25)$	12	0	88	4.380
	CV$d(d = 30)$	8	7	85	7.083
	Adaptive CCP	5	0	95	4.389
$\beta = (2,9,6,4,8)'$	full model	0	0	100	5.231
	C_p	0	0	100	5.231
	BIC$(\lambda = \log n)$	0	0	100	5.231
	GIC$(\lambda = n/\log n)$	0	6	94	6.608
	CV$d(d = 1)$	0	8	92	7.849
	CV$d(d = 20)$	0	1	99	5.441
	CV$d(d = 25)$	0	6	94	6.843
	CV$d(d = 30)$	0	40	60	18.627
	Adaptive CCP	0	0	100	5.231
$\beta = (1,2,3,2,3)'$	full model	0	0	100	5.169
	C_p	0	68	32	5.717
	BIC$(\lambda = \log n)$	0	85	15	6.238
	GIC$(\lambda = n/\log n)$	0	100	0	13.689
	CV$d(d = 1)$	0	63	37	8.728
	CV$d(d = 20)$	0	84	16	8.616
	CV$d(d = 25)$	0	93	7	12.114
	CV$d(d = 30)$	0	100	0	20.488
	Adaptive CCP	0	85	15	6.748

the hopes that researchers like Jun Shao will continue to apply their considerable talents to shed light on this important area.

Department of Biostatistics, Cleveland Clinic Foundation.

Department of Preventive Medicine and Biostatistics, University of Toronto, Toronto, Ontario M5S 1A8, Canada.

COMMENT

Mervyn Stone

University College London

This paper skilfully clarifies a number of important questions concerning model selection for least-squares prediction with squared error loss. It also suggests some problems that have not yet been decently formulated.

Professor Shao has responsibly concentrated on sorting out the logic and technical kernel of some necessary mathematics. I will exercise discussant's license to range irresponsibly over the wider framework where intuition and conjecture can be countenanced.

(1). The paper focuses on selection within a given set of linear models: it does not consider the question of choice in the 'given'— a possibility that must be sensitive to scientific context. In my 1974 paper, I considered cross-validatory choice of a linear model for the rotation-averaged shape of the Earth, for which there can be no scientific limit to p_n— the number of Legendre polynomials in what must regarded as a "soft science" approximation to the true shape. But in another example of earthy statistics—geodetic survey—the appropriate linear model is unambiguously 'given' by the topology of the triangulation points (with, incidentally, no room for further selection *within* the model). I think the first case — of potentially unlimited p_n—is more representative of statistical practice with linear models than is the case where p_n is 'given'. All of which raises the following lurking questions when we do fix p_n: Could the least-squares measure $L_n(\alpha_n^L)$ (whose value is crucial in the paper's consideration of asymptotic validity) be appreciably reduced by enlarging the model? — and, even if it were not, should we not admit such extension by considering other predictors that avoid the least-squares pitfall of over-parametrization. Of course, the asymptotics may then prove unmanageable!

(2). Selection for prediction is not as censorious, about not picking the minimal correct model, as any method that wants to pin down the truth in some "hard science". Shao's use of the loss ratio (2.3), as in Li (1987), sets the mathematics in the permissive direction (whereas his tables of selection probabilities seem tangential to the main message of the paper). But it is not clear, from the theorems proved, whether this is a "distinction without a difference" i.e. whether, in those cases where a method has asymptotic validity (in the sense of (2.3)), it does this only by asymptotically selecting the minimal correct model (if there is one)—ruling out any trade-off between 'bias' (in Δ_n) and 'variance' represented by the second term in L_n.

(3). I would like to have a clearer intuition about the necessity status of condition (2.6). Li referred to his stronger version of this condition as "reasonable". My rough verbalisation of (2.6) is: "For every incorrect model α, either $p_n(\alpha) \to \infty$ or, if $\sqrt{\Delta_n(\alpha)} \to 0$, it does so *slower* than $1/\sqrt{n}$—with a further combined condition on the rates of these limits when the number of incorrect models goes to infinity". (The square roots keep things on the observation scale.) That (with reference to Theorem 1(i)) things would get worse for $\hat{\alpha}_{n,2}$ if we were to make the incorrect models get nearer the truth faster then $1/\sqrt{n}$ offends my current intuition. But I am open to persuasion. In his challenging 1988 paper—for the case of p_n fixed, more than one true model, and a criterion given by the probability of selection of the minimal correct model—Shao imposed the condition that $\Delta_n(\alpha)$ be bounded away from zero for incorrect models, on the reasonable grounds that this was an identifiability condition "very minimal for asymptotic analysis". Its present relaxation could, I think, have been taken further.

(4). The findings for delete-1 and delete-d crossvalidatory choice are most interesting. In 1973, I conjectured that "delete-1" would be superior—on the grounds that the n delete-1 predictors came closer to the whole-sample predictor and that their assessment still used the whole sample. Shao's (1988) paper made the "shocking discovery" that, under the conditions he imposed, rectification of the inconsistency of delete-1 crossvalidatory choice (with respect to the probability of selection of the minimal correct model) required that $d/n \to 1$! This result has been picked up and risks becoming a stable myth of the form "Delete-d (large) Good, Delete-1 Bad.". For example, chemometrician Clementi (1995) refers to Shao (1988) in this unconditional remark: "*statisticians agree that group formation* [i.e. delete-d in some pattern] *is better that LOO* [leave one out] *one theoretical grounds*". I hope that both statisticians and practitioners will read this new work of Shao's which gives a well-balanced overview of the present position about the asymptotics.

(5). As far as I can see, Shao's work has revealed something of a cleavage between what the asymptotics say about prediction and the idea that we should in many problems be able to benefit from some trade-off between bias and variance. Perhaps the clear-cut *selection* of a linear model (setting some parameters to zero that require only Tukey's *flattening* or Stein's *shrinkage*) goes too far for the trade-off to show itself (except to a small extent with the designedly variance sensitive CV_d in Tables 1 and 2). Or is there a need for alternative asymptotics?

(6). I have already indicated why I think further work is required for the soft science case where interest lies in predictive efficiency. The problem is that, with models α indexed by three controlling parameters n, $p_n(\alpha)$, $\Delta_n(\alpha)$, there are many ways of going to infinity (even for the problem in probability of limit laws for sums of i.i.d. random variables, there is a spectrum that extends from central limit theorems to those for moderate and large deviations). The further problem for the practitioner is to know which of these different routes to infinity is the one that will offer guidance for the finite problem in hand.

25 Hawtrey Drive, Ruislip Middx, London HA4 8QW, UK.

COMMENT

Ping Zhang

University of Pennsylvania

1. Introduction

Model selection is a difficult problem for two reasons: First, related to the problem are fundamental philosophical issues such as the existence of a true model and the ultimate goal of statistical modeling. Second, the topic is so broad that precise definition of the model selection problem seems both technically implausible and practically unnecessary. Ironically, applications of model selection, especially linear model selection, are ubiquitous in many areas of empirical research. To some extent, we could even argue that most statistical problems, from hypothesis testing to nonparametric function estimation, are related to the idea of model selection. The potential scope of a general model selection problem therefore goes far beyond variable selection in linear regression, which is the subject treated in the paper under discussion. Despite the limitation in scope, Professor Shao's rigorous treatment of the subject has not only unified and improved many existing results, but also clarified misconceptions and brought new insights into the behavior of a large class of model section methods. The role of model dimension, to my knowledge, is previously not well understood (c.f., Example 2). Professor Shao is to be commended for taking on such a difficult subject. The work under discussion is a much needed service to the statistical community. There is no doubt in this investigator's mind that Professor Shao's work will soon become a standard reference in the model selection literature.

2. AIC vs. BIC

As statisticians, we should always bear in mind that mathematical results do not automatically render themselves statistical interpretations. The main conclusion in Shao's paper is that GIC_{λ_n} with $\lambda_n = 2$ and $\lambda_n \to \infty$ (which I shall refer to as AIC-like and BIC-like criteria respectively) represent two classes of model selection criteria whose asymptotic behavior are fundamentally different. The validity of each class is associated with the structures of the unknown true model. An implicit assumption in this argument is therefore the existence of a true model. According to Shao's results, BIC-like criteria would perform better if the true model has a simple structure (finite-dimension) and AIC-like criteria would do better if the true model is a complex one (infinite-dimension). The results are undisputable so far as the mathematics is concerned. In practice, however, there is a flip side to this interpretation. An argument can be made in favor of BIC-like criteria regardless of the true model. First of all, one should realize that statistical models are mostly used in areas where the existence of a true model is doubtful. Even if a true model does exist, there is still ample reason to choose simplicity over correctness knowing perfectly well that the selected model might be untrue. The practical advantage of a parsimonious model often overshadows concerns over the correctness of the model. After all, the goal of statistical analysis is to extract information rather than to identify the true model. In other words, the parsimony principle should be applied not only to candidate fit models, but the true model as well. Theoretically, this is in line with the argument of Rissanen (1986b), where a BIC-like criterion is shown to be optimal from an information theoretic point of view.

3. The Role of Loss Function

The point, of course, is that optimality (e.g., loss-efficiency) is a concept that depends on the objective. Shao argues repeatedly that the GIC_{λ_n} criterion with $2 < \lambda_n < \infty$ does not merit further attention because the asymptotic properties of the corresponding GIC_{λ_n} criterion is dominated by either $\lambda_n = 2$ or $\lambda_n \to \infty$. This, however, is the result of using $L_n(\alpha)$ as an all purpose loss function. The Γ_{n,λ_n} criterion can be viewed as a sample estimate of $L_n(\alpha)$ if and only if $\lambda_n = 2$. Intuitively, it is unfair to measure the performance of other GIC_{λ_n} criteria using a loss function that is derived mainly for the case of $\lambda_n = 2$. In fact, if one is willing to modify the loss function, it is possible to show that any $2 < \lambda_n < \infty$ is loss-efficient. To see this, let us define a new loss function

$$\tilde{L}_n(\alpha) = \frac{1}{dN} \sum_{s \in \mathcal{S}} \|\boldsymbol{\mu}_{n,s} - \hat{\boldsymbol{\mu}}_{n,s}(\alpha)\|^2, \tag{1}$$

where the notations are the same as in Section 4 of Shao's paper. The loss function in (1) is equivalent to the conditional prediction error of predicting d

future data points with $n-d$ current data points. The delete-d CV criterion, i.e., $\mathrm{CV}_{n,d}(\alpha)$ in Shao's notation, is a sample estimate of $\tilde{L}_n(\alpha)$. It is natural to expect, and I would be surprised if it is not the case, that $\mathrm{CV}_{n,d}(\alpha)$ is loss-efficient under the modified loss function $\tilde{L}_n(\alpha)$. Suppose that this were true. Then following Shao's argument, the GIC_{λ_n} criterion with $\lambda_n = n/(n-d)+1$ can be shown to be loss efficient if one uses $\tilde{L}_n(\alpha)$ as the loss function. Notice that $2 < \lambda_n < \infty$ if d is chosen to be proportional to n.

Finally, if one uses the accumulated prediction error of Wei (1992) to replace $L_n(\alpha)$, then, contrary to Shao's conclusion, GIC_{λ_n} with $\lambda_n = \log(n)$ can be shown to be loss-efficient. The moral here is that one should not take theory out of its context. The choice of loss function has a tremendous bearing on the asymptotic properties of the corresponding model selection criterion. Occasionally, casual interpretation and generalization of theoretical results can be misleading to the novice reader. I am basically in agreement with Shao regarding the distinction between $\lambda_n = 2$ and $\lambda_n \to \infty$, except for minor differences in the interpretation of results. However, I disagree with Shao's claim that the case $2 < \lambda_n < \infty$ is uninteresting (see Zhang (1992)). What we have demonstrated in the previous paragraph is that every member of the GIC_{λ_n} class with $2 < \lambda_n < \infty$ can be asymptotically optimal, provided that we define optimality properly. Likewise, we should be able to differentiate and to justify different GIC_{λ_n} criteria when $\lambda_n \to \infty$ at different rates. This latter problem, however, is rarely discussed in the literature.

4. Extensions

The challenge of the model selection problem is that, without assuming the existence of a true model, it is rather difficult to assess the merit of a proposed method objectively. Each method has some merit in its own right. For example, the Bayesian approach has the philosophical advantage that one is not forced to choose a single model out of a set of possible models. The classical argument of Akaike (1973) states that the best model should be the one that yields the highest predictive power. Rissanen (1986c) asserts that the best model should be the simplest one that is capable of fully describing the data. These different approaches seem to have nothing in common. The general consensus is, however, that most of the existing model selection criteria give rise to a quantification of the parsimony principle. They differ in their capacity to balance goodness-of-fit and model complexity.

In Shao's work, the GIC_{λ_n} criterion is used as a prototype class of model selection procedures that, when λ_n varies, represents different levels of trade-off between goodness-of-fit and model complexity. A natural extension of GIC, in Shao's notation, is

$$\mathrm{GIC}^*(\alpha) = S_n(\alpha) + \lambda_n \hat{\sigma}_n^2 a(p_n(\alpha)), \tag{2}$$

where $a(p)$ is an arbitrary function of p. The GIC_{λ_n} criterion corresponds to $a(p) = p$. George and Foster (1994) show that the GIC^* criterion in (2) with $a(p) = \log(p)$ has some minimax property under a newly defined loss function. Hartigan (personal communication) finds that GIC^* with $a(p) = p^2$ has a similar interpretation. Still more general extensions of GIC_{λ_n} can also be found in the literature. For example, Wei (1992) proposes what he calls an FIC criterion

$$\mathrm{FIC}(\alpha) = \hat{\sigma}_n^2 \left\{ n + \log\det(\mathbf{X}_n(\alpha)\mathbf{X}_n'(\alpha)) \right\}, \tag{3}$$

where $\sigma^{-2}\mathbf{X}_n(\alpha)\mathbf{X}_n'(\alpha)$ is the Fisher information matrix under the model indexed by α. Note that (3) cannot be written in the form of (2) since the penalty term in (3) is not necessarily a function of model dimension $p_n(\alpha)$. As a general theory, it would be nice if the current results in Shao's paper can be extended to criteria such as (2) and (3).

5. Linear Models for Panel Data

Panel data, i.e., longitudinal records taken from a randomly selected group of panelists, arise frequently in econometrics and other social sciences (Hsiao (1986)). The general format of the observed data is $(\mathbf{x}_i, \mathbf{z}_t, y_{it}), i = 1, \ldots, N; t = 1, \ldots, T$, where $\mathbf{x}_i$ is a vector of independent variables measuring the demographic attributes of the ith panelist; $\mathbf{z}_t$ is a vector of variables that measures changes in the environment; and y_{it} is a response variable observed at time t from the ith panelist. What sets panel data apart from conventional data is that the observations vary not only across individuals (as in cross-sectional survey data), but also across time (as in aggregate time series records). Different types of models are often needed to describe the two types of variation. Suppose that we fit a linear model of the following form:

$$y_{it} = \mu + \mathbf{x}_i\theta + \mathbf{z}_t\beta + \epsilon_{it}, \tag{4}$$

where ϵ_{it} are i.i.d. $(0, \sigma^2)$ across both i and t. Apparently, variable selection under model (4) can be accomplished by using any of the criteria established for ordinary regression models. A closer look at the situation suggests, however, that conventional model selection methods may not be appropriate for the purpose of panel data analysis.

Take the predictive approach for example. For panel data, it is often more relevant to predict aggregate statistics rather than individual values of future observations. Let $\mathbf{g}_t$ denote a summary statistics of $y_{it}, i = 1, \ldots, N$. Let $\hat{\mathbf{g}}_t$ be a predictor of $\mathbf{g}_t$ based on data up to time $t-1$. Define the accumulated prediction error as

$$\mathrm{APE} \approx \sum_{t=t_0}^{T} \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2.$$

Under ordinary regression models (i.e., $N = 1$), it has been shown that APE with least squares predictors is asymptotically equivalent to the BIC criterion (Wei (1992)). Hence the results in Shao's paper apply to APE as well. For panel data (i.e., $N > 1$), depending on the target function $\mathbf{g}_t$, the asymptotic properties of APE can be rather different.

For simplicity, suppose that the covariates $\mathbf{x}_i$ and $\mathbf{z}_t$ are iid random vectors following the standard multivariate normal distribution. When $\mathbf{g}_t$ is the cross-sectional sample mean, i.e., $\mathbf{g}_t = \bar{y}_t = N^{-1}\sum_{i=1}^{N} y_{it}$, Zhang (1996) shows that

$$\mathrm{APE} \approx N^{-1}\Delta_1(T) + \sigma^2 \dim(\beta) N^{-1} \log(T), \tag{5}$$

where $\Delta_1(T)$ is an approximate measure of goodness-of-fit. An interesting observation is that (5) does not penalize cross-sectional model complexity since $\dim(\theta)$ does not appear in the expression. More to the point, this case is likely to be covered by Professor Shao's theory because (5) is, qualitatively speaking, a member of the GIC_{λ_n} class.

Next, suppose that we wish to predict the cross-sectional variance, i.e., $\mathbf{g}_t = N^{-1}\sum_{i=1}^{N}(y_{it} - \bar{y}_t)^2$. A result of Zhang (1996) implies that

$$\mathrm{APE} \approx N^{-1}\Delta_2(T) + 4\sigma^2(\sigma^2 + \|\theta\|^2) N^{-1} \log(T), \tag{6}$$

where Δ_2, as before, is a measure of goodness-of-fit. Contrary to the previous case, we note that (6) does not penalize cross-time complexity. Furthermore, (6) is not a member of the GIC_{λ_n} class and Shao's results do not apply. The result for general $\mathbf{g}_t$ is more complicated (see Zhang (1996)).

In the past two sections, we have demonstrated that some important model selection criteria do not fit into the framework of Shao's paper. Our purpose is not to show that there are pathological exceptions to an otherwise nice theory. Instead, we believe that Professor Shao's results can be extended to much broader contexts. The key here, as we pointed out at the beginning, is to establish a general definition of what a model selection problem is and what one means by optimality.

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A.

REJOINDER

Jun Shao

I would like to thank the Editor and an Associate Editor for organizing this discussion. I am also grateful to discussants for providing very insightful discussions, useful additional results, and directions for further research. As the discussants pointed out, model selection is a very complicate and difficult problem. I hope that more discussions of this kind will be seen in the future and that the current paper will serve as a starter for theoretical research in assessing various existing model selection procedures and in developing new methods. In the following I focus on some major issues raised in the discussion, instead of replying to each discussant separately. I shall adopt the same notation that I used in the main paper.

1. Criteria of Assessing Model Selection Procedures

Any theoretical study in assessing some statistical procedures must be based on one or several criteria. For example, the most commonly used criteria in an estimation problem include the mean squared error (or, more generally, the risk), consistency, asymptotic efficiency, admissibility, etc. Model selection is far more complicated than an estimation problem and using a single criterion may not be sufficient in many situations.

I stated in Section 1 that the goal of model selection is to minimize the squared error loss $L_n(\alpha) = \|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha)\|^2/n$ over models $\alpha \in \mathcal{A}_n$. Ideally, the loss function L_n should be used to assess model selection procedures. But using L_n (or the mean squared error) to assess model selection procedures is very difficult or impossible. Criterion (2.3) (called asymptotic loss efficiency) guarantees that when the sample size n is large, $L_n(\hat{\alpha}_n)$ is close to $\min_\alpha L_n(\alpha)$ (unfortunately, we do not know how large is large, which is a limitation of all asymptotic analysis). It is a good starting point for theoretical research in this area, although many other issues need to be worried about. The scenario is very similar to an estimation problem in which one is not able to assess the finite-sample mean squared error but considers consistency and asymptotic efficiency instead. Consistency is an asymptotic analog of admissibility in the sense that we should not encourage the use of inconsistent (inadmissible) procedures unless there are specific reasons, but usually there are many consistent (admissible) procedures that have to be further assessed.

Rao-Tibshirani wondered why I also consider criterion (2.1), the consistency. Criteria (2.1) and (2.3) are related and are equivalent in some cases (Proposition 1). Criterion (2.3) focuses on the loss $L_n(\hat{\alpha}_n)$, whereas criterion (2.1) emphasizes the frequency of $\hat{\alpha}_n = \alpha_n^L$, where $\hat{\alpha}_n$ is the model selected using a model selection procedure and α_n^L is the optimal model that minimizes $L_n(\alpha)$. If α_n^L is non-random, then criterion (2.1) is the same as the consistency in an estimation problem where $\hat{\alpha}_n$ is viewed as an estimator of α_n^L. In some problems (e.g., Examples 1 and 2), the optimal model can be defined as the correct model with the smallest dimension and, therefore, criterion (2.1) does not depend on any loss function. On the other hand, criterion (2.3) is loss-dependent, which will be further discussed later. For these reasons, it is of interest to study model selection procedures under both criteria (2.1) and (2.3).

For a fixed sample size n, criteria (2.1) and (2.3) are different as pointed out by *Rao-Tibshirani*, but I think that a model selection procedure having a high frequency of choosing the optimal model should usually be fairly good in terms of the loss L_n. I am very grateful to *Rao-Tibshirani* for their simulation results that complements my simulation study. From their Table 1, one can find that except for the case where the full model is the only correct model, the GIC with $\lambda = n/\log n$ and the delete-25 CV perform fairly well in terms of the average loss L_n, although they may not be the best.

Theoretical research based on other criteria is called for. *Beran* adopted an asymptotic minimax criterion. Perhaps we may consider the convergence rates of consistent model selection procedures. For example, if we can show that

$$1 - \frac{\min_\alpha L_n(\alpha)}{L_n(\hat{\alpha}_n^A)} = O_p(a_n) \qquad \text{and} \qquad 1 - \frac{\min_\alpha L_n(\alpha)}{L_n(\hat{\alpha}_n^B)} = O_p(b_n),$$

where $\hat{\alpha}_n^A$ and $\hat{\alpha}_n^B$ are models selected by model section procedures A and B, respectively, and a_n and b_n are two positive sequences of numbers satisfying $a_n/b_n \to 0$, then procedure A is better than procedure B.

2. Loss Function

Criterion (2.3) depends on the loss function

$$L_n(\alpha) = \Delta_n(\alpha) + \frac{\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha) \boldsymbol{e}_n}{n},$$

where $\Delta_n(\alpha)$ is a squared "bias" term and $\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha) \boldsymbol{e}_n/n$ is a "variance" term related to the complexity of model α. *Zhang* raises the question of using a loss function that puts heavier penalty on the complexity of models. Indeed, we may consider the following loss function

$$\tilde{L}_n(\alpha) = \Delta_n(\alpha) + \frac{(\lambda_n - 1)\boldsymbol{e}_n' \boldsymbol{H}_n(\alpha) \boldsymbol{e}_n}{n},$$

which is equivalent to the loss function in *Zhang* with certain choice of λ_n. Note that $\lambda_n - 1$ can be viewed as a penalty parameter on the complexity of a model and that $L_n(\alpha)$ is simply $\tilde{L}_n(\alpha)$ with $\lambda_n \equiv 2$. The cost-complexity criterion in *Rao-Tibshirani* with a fixed λ is also closely related to the loss function $\tilde{L}_n$.

Theorem 6. *Assume that* (2.6) *and* (3.4) *hold and that* $\hat{\sigma}_n^2$ *is consistent for* σ^2. (i) *If* $\lambda_n \to \infty$, *then the* GIC_{λ_n} *is asymptotically* $\tilde{L}_n$*-loss efficient, i.e.,*

$$\frac{\tilde{L}_n(\hat{\alpha}_{n,\lambda_n})}{\min_\alpha \tilde{L}(\alpha)} \to_p 1.$$

(ii) *The same conclusion holds for* $\lambda_n \equiv \lambda$ (*a fixed constant*), *provided that* $\mathcal{A}_n$ *contains at most one correct model for all* n.

The proof of this result is given in the end.

Theorem 6(ii) indicates that if we consider the loss function $\tilde{L}_n$ with $\lambda_n \equiv \lambda$, then the GIC_{λ_n} with $\lambda_n \equiv \lambda > 2$, which is also known as the FPE_λ method, plays the same role as the C_p method in the case where L_n is used as the loss function. But the FPE_λ is still not asymptotically loss efficient if $\mathcal{A}_n$ contains more than one correct models.

Theorem 6(i) indicates that the GIC_{λ_n} with $\lambda_n \to \infty$ is asymptotically $\tilde{L}_n$-loss efficient, which is natural since the same λ_n is used in the loss function and the GIC. But why is the GIC_{λ_n} with $\lambda_n \to \infty$ also asymptotically loss efficient when the squared error loss L_n is used (Theorem 2)? The following result answers this question.

Theorem 7. (i) *If there exists a fixed-dimension correct model* (*in* $\mathcal{A}_n$ *or not in* $\mathcal{A}_n$) *and* $\lambda_n/n \to 0$, *then asymptotic* L_n*-loss efficiency is the same as asymptotic* $\tilde{L}_n$*-loss efficiency, i.e.,*

$$P\left\{\alpha_n^L = \alpha_n^{\tilde{L}} \textit{ for sufficiently large } n\right\} = 1,$$

where α_n^L *and* $\alpha_n^{\tilde{L}}$ *are the optimal models under the loss functions* L_n *and* $\tilde{L}_n$, *respectively.*
(ii) *If* $\mathcal{A}_n^c$ *is empty for all* n *and*

$$\frac{(\lambda_n\hat{\sigma}_n^2 - 2\sigma^2)p_n(\alpha_n^L)}{nL_n(\alpha_n^L)} \to_p 0,$$

then

$$\frac{L_n(\alpha_n^L)}{L_n(\alpha_n^{\tilde{L}})} \to_p 1 \qquad \textit{and} \qquad \frac{\tilde{L}_n(\alpha_n^L)}{\tilde{L}_n(\alpha_n^{\tilde{L}})} \to_p 1.$$

Thus, if there is a fixed-dimension correct model, asymptotic loss efficiency with different loss functions $\tilde{L}_n$ are equivalent and the GIC_{λ_n} with $\lambda_n \to \infty$ is asymptotically loss efficient regardless of which loss function is used. To respond to *Zhang*'s comment on the FPE_λ with $2 < \lambda < \infty$, I find that my comment (in the end of Section 3) on the relative performance between the FPE_λ and the GIC_{λ_n} is valid even if L_n is replaced by $\tilde{L}_n$; however, my comment on the relative performance between the FPE_λ and the GIC_2 applies only to the case where the squared error loss is used.

3. Signal-to-Noise Ratio

Several discussants addressed the problem of signal-to-noise ratio. When n is fixed and the ratio $\|\boldsymbol{\beta}\|/\sigma \to 0$, all model selection procedures will fall apart, since we cannot distinguish the zero components and the non-zero components of $\boldsymbol{\beta}$. The GIC_{λ_n} with a large λ_n is more sensitive to the signal-to-noise ratio than the GIC_{λ_n} with a small λ_n, which is numerically illustrated by the last two cases in Table 1 of *Rao-Tibshirani.* What can we do when $\|\boldsymbol{\beta}\|/\sigma$ is very small? Perhaps a different asymptotic framework should be adopted as *Stone* suggested. We may consider different convergence or divergence rates of n, $p_n(\alpha)$, and $\Delta_n(\alpha)$ to provide an asymptotic analysis that can offer the best guidance for a given practical situation.

4. GIC_2 versus GIC_{λ_n}

Beran, *Stone*, and *Zhang* discussed the choice between the GIC_2 (or the C_p) and the GIC_{λ_n} (or the choice between the delete-1 CV and the delete-d CV). Asymptotic minimaxity cannot distinguish these two methods when $r = 0$ (*Beran*). Under criteria (2.1), (2.3) and the asymptotic settings considered in the current paper, the GIC_{λ_n} with a large λ_n is preferred, because the only situation where the GIC_2 is possibly better is when there is no fixed-dimension correct model *and* the squared error loss L_n is used. As I discussed above, however, the GIC_2 is less sensitive to the small signal-to-noise ratio, although it also breaks down as $\|\boldsymbol{\beta}\|/\sigma \to 0$. We may need to consider other criteria in making a choice.

5. Choices of λ_n

Even when we decide to adopt the GIC_{λ_n} with $\lambda_n \to \infty$, we still need to choose a particular λ_n. This issue is not addressed in the current paper and seems to be a difficult problem. A promising adaptive method is introduced by *Rao-Tibshirani.* Their method amounts to finding a suitable λ_n by minimizing an objective function via methods such as the cross-validation. We may even use this method to assess the GIC_2 and the GIC_{λ_n}. Properties of this method need to be investigated.

6. Condition (2.6)

Stone questioned condition (2.6) in his comment (3). My first reaction is that (2.6) is a weak condition if we focus on the case where $n \to \infty$ and $\|\boldsymbol{\beta}\|/\sigma$

is fixed. As I discussed in the end of Section 2, in several important cases (2.6) is the same as (2.7): $n\Delta_n(\alpha) + \sigma^2 p_n(\alpha) \to \infty$ for all $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$. If $p_n(\alpha)$ is bounded, then very likely $\Delta_n(\alpha)$ is bounded away from 0. My second reaction to *Stone*'s comment (3) is that if $\Delta_n(\alpha)$ tends to 0 faster than n^{-1}, then it is hard to distinguish models and all model selection procedures may break down (it hurts the GIC_{λ_n} with a large λ_n more, I believe). This is very similar to the situation where we fix n and let $\|\boldsymbol{\beta}\|/\sigma$ tend to 0.

7. Other Research Problems

The discussants pointed out various other directions for research in this area. For example, *Zhang*'s extended GIC; *Rao-Tibshirani*'s adaptive method of choosing λ; *Zhang*'s extension of model selection to panel data; the use of different asymptotic settings (*Stone*); the choice between the C_p and the GIC_{λ_n} (*Beran*, *Stone*, and *Zhang*); the use of shrinkage estimators (*Beran* and *Stone*), etc. In some of these problems we may need to start with empirical studies. I hope that more researchers will work on these problems that are pertinent to applications of model selection.

8. Proofs

Proof of Theorem 6. Using the conditions of the theorem we can establish

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + \tilde{L}_n(\alpha) + o_p\left(\tilde{L}_n(\alpha)\right),$$

where the o_p is uniformly in $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$ and is uniformly in $\alpha \in \mathcal{A}_n$ if $\lambda_n \to \infty$. Then the results in (i) and (ii) follow from

$$0 \le \frac{\Gamma_{n,\lambda_n}(\alpha_n^{\tilde{L}}) - \Gamma_{n,\lambda_n}(\hat{\alpha}_{n,\lambda_n})}{\tilde{L}_n(\hat{\alpha}_{n,\lambda_n})} = \frac{\tilde{L}_n(\alpha_n^{\tilde{L}}) - \tilde{L}_n(\hat{\alpha}_{n,\lambda_n})}{\tilde{L}_n(\hat{\alpha}_{n,\lambda_n})} + o_p(1) \le o_p(1),$$

where the inequalities follow from $\Gamma_{n,\lambda_n}(\alpha_n^{\tilde{L}}) \ge \Gamma_{n,\lambda_n}(\hat{\alpha}_{n,\lambda_n})$ and $\tilde{L}_n(\alpha_n^{\tilde{L}}) \le \tilde{L}_n(\hat{\alpha}_{n,\lambda_n})$.

Proof of Theorem 7. If there is a fixed-dimension correct model, then $\liminf_n \min_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c} \Delta_n(\alpha) > 0$ (see Nishii (1984)). Hence the result in (i) follows from the fact that for sufficiently large n, both α_n^L and $\alpha_n^{\tilde{L}}$ are the same as the model that minimizes $\Delta_n(\alpha)$, $\alpha \in \mathcal{A}_n$, and has the smallest dimension. Result (ii) follows from

$$\begin{aligned} 0 \le \frac{\tilde{L}_n(\alpha_n^L) - \tilde{L}_n(\alpha_n^{\tilde{L}})}{\tilde{L}_n(\alpha_n^L)} &\le \frac{L_n(\alpha_n^L) - L_n(\alpha_n^{\tilde{L}})}{L_n(\alpha_n^L)} + \frac{(\lambda_n\hat{\sigma}_n^2 - 2\sigma^2)[p_n(\alpha_n^L) - p_n(\alpha_n^{\tilde{L}})]}{nL_n(\alpha_n^L)} \\ &\le \frac{(\lambda_n\hat{\sigma}_n^2 - 2\sigma^2)[p_n(\alpha_n^L) - p_n(\alpha_n^{\tilde{L}})]}{nL_n(\alpha_n^L)} = o_p(1). \end{aligned}$$

Additional References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiado, Budapest.

Beran, R. and Dümbgen, L. (1996). Modulation estimators and confidence sets. *Beiträge zur Statistik* **31**. Institut für Angewandte Mathematik, Universität Heidelberg.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.

Clementi, S. (1995). Comment: Data analyses or problem formulations? *J. Chemometrics* **9**, 226-228.

George, E. I. and Foster, D. P. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge University Press, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. on Math. Statist. Prob.* **1** (Edited by J. Neyman), 361-380. University of California Press, Berkeley.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16**, 120-133.

Rao, J. S. (1994). Adaptive subset selection via cost-optimization using resampling methods in linear regression models. Ph.D. dissertation, University of Toronto.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Rissanen, J. (1986b). Order estimation by accumulated prediction errors. *J. Appl. Probab.* **23**, 55-61.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. on Math. Statist. Prob.* **1** (Edited by J. Neyman), 107-206. University of California Press, Berkeley.

Zhang, P. (1992). On the distributional properties of model selection criteria. *J. Amer. Statist. Assoc.* **87**, 732-737.

Zhang, P. (1996). APE and model selection: Linear models for panel data. Technical report, Department of Statistics, University of Pennsylvania.